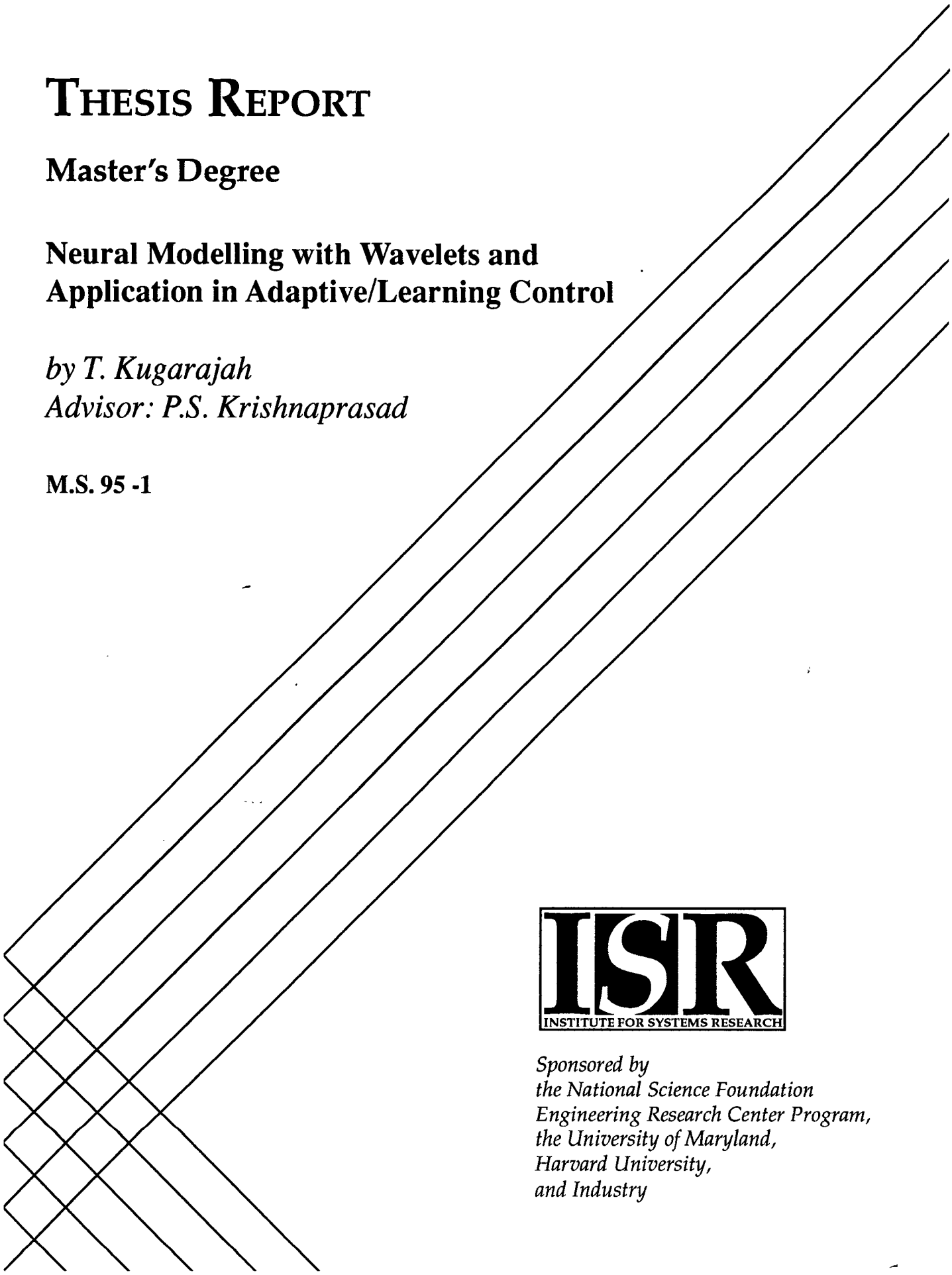
THESIS REPORT

Master's Degree

Neural Modelling with Wavelets and Application in Adaptive/Learning Control

by T. Kugarajah
Advisor: P.S. Krishnaprasad

M.S. 95 -1



Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1995		**00-00-1995 to 00-00-1995**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Neural Modelling with Wavelets and Application in Adaptive/Learning Control	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,The Graduate School,2123 Lee Building,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**91**	

Abstract

Title of Thesis: Neural Modelling with Wavelets and Application in Adaptive/Learning Control

Name of degree candidate: Tharmarajah Kugarajah

Degree and year: Master of Science, 1995

Thesis directed by: Professor P. S. Krishnaprasad
Department of Electrical Engineering
and
Associate Professor W. P. Dayawansa
Department of Electrical Engineering

Spatio-spectral properties of the Wavelet Transform provide a useful theoretical framework to investigate the structure of neural networks. A few researchers (Pati & Krishnaprasad, Zhang & Benveniste) have investigated the connection between neural networks and wavelet transforms. However, a number of issues remain unresolved especially when the connection is considered in the multi-dimensional case. In our work, we resolve these issues by extensions based on some theorems of Daubechies related to wavelet *frames* and provide a framework to analyze *local learning* in neural-networks.

We also provide a constructive procedure to build networks based on wavelet theory. Moreover, cognizant of the problems usually encountered in practical implementations of these ideas, we develop a heuristic methodology, inspired by similar work in the area of Radial Basis Function (RBF) networks (Moody & Darken, Platt), to build a network sequentially *on-line* as well as off-line.

We show some connections of our method to some existing methods such as Projection Pursuit Regression (Friedman), Hyper Basis Functions (Poggio & Girosi) and other methods that have been proposed in the literature on neural-networks as well as statistics. In particular, some classes of wavelets can also be derived from the regularization theoretical framework given by Poggio & Girosi.

Finally, we choose *direct nonlinear* adaptive control to demonstrate the utility of the network in the context of *local learning.* Stability analysis is carried out within a standard Lyapunov formulation. Simulation studies show the effectiveness of these methods. We compare and contrast these methods with some recent results obtained by other researchers using Back Propagation (Feed-Forward) Networks, and Gaussian Networks.

Neural Modelling with Wavelets and Application in Adaptive/Learning Control

by

Tharmarajah Kugarajah

Thesis submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Master of Science
1995

Advisory Committee:

Professor P. S. Krishnaprasad, Chairman/Advisor
Associate Professor W.P. Dayawansa, Co-advisor
Associate Professor S. Shamma

Dedication

To my family

Acknowledgements

I would like to thank my academic advisor Professor Krishnaprasad for his guidance and encouragement. His knowledge of diverse areas, his willingness to explore unchartered territory, along with the freedom he granted, inspired me to pursue many diverse topics. I am also indebted to Prof. W.P. Dayawansa, my co-advisor, who provided useful comments and criticisms that enhanced my understanding of various mathematical ideas. He also helped me in various non-academic problems, and I can never adequately thank him. I also thank Dr. Qinghua Zhang of IRISA-INRIA, France, for useful interaction on the material in Chapter 2 of this thesis. Thanks are due to Dr. S. Shamma for agreeing to be a member of my thesis committee.

Several of my colleagues at the Institute for Systems Research, and in particular Vikram and Dimitris at the Intelligent Servosystems Laboratory, provided a supportive environment and useful tips on Word-Processing, Programming, etc. I owe thanks to all of them.

I also thank my family for all the support throughout the years. Finally, I must thank several people in Sri Lanka– teachers, professors and friends at Peradeniya–too numerous to mention individually, from whom I benefited over the years.

Financial support for this research was provided in part by a graduate school fellowship from the University of Maryland, and in part by the Institute for

Systems Research under National Science Foundation's Engineering Research Centers Program: NSFD CDR 8803012, and by the AFOSR Grant AFOSR-F49620-92-J-0500, and by the Army Research Office under Smart Structures URI Contract No. DAAL03-92-G-0121.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

'Learning' can be viewed as providing an approximation to a desired mapping within a given tolerance. From an electrical engineering perspective, interest in theories of learning arises owing to the presence of many systems that are unknown or only partially known and are difficult to model. In such situations the mapping needs to be implemented from observations during interactions with the system. To solve this problem, researchers in several disciplines have developed tools that can be graphically interpreted as 'networks'. Although these tools initially derived some inspiration from biological observations, approximation theory and statistical/information-theoretic methods have been recognized as essential tools to tackle the enormous complexity inherent in the method. Reflecting this diversity of disciplines, and depending on the application domain, such networks are often known variously as 'neural networks', 'statistical networks', 'connectionist networks' and 'biological networks'.

In spite of the explosive growth of research in this area in recent years, the methodology has largely remained heuristic; precise mathematical methods are often difficult to derive or when derived, remain largely without any practical

merit. As a result, tools that can provide more insight into their structure and 'de-mystify' them are important. One finds such a tool in wavelets and in this thesis we focus our attention on how well this tool can help in this task.

First we have to address the issue of learning from the viewpoint of approximation theory. For this purpose, we use the theory of wavelets with wavelet as an alternative to implementing *local* learning with Gaussian Radial Basis Functions(GRBF). In a related spirit, Poggio and Girosi [27] showed that Radial Basis Function Networks can be derived from Regularization Theory and Pati and Krishnaprasad [25, 23] have shown that feed-forward neural networks can be considered within the framework provided by discrete wavelet transform theory. Zhang and Benveniste [34] give a somewhat different treatment of this connection between neural networks and wavelet transforms. However, adequate treatment of theoretical issues (e.g., the construction of wavelet frames, method of dilating, bounds on error in approximation using a finite subset of dilations and translations) in high dimensional problems, or practically feasible ways to tackle the ubiquitous problem of 'curse of dimensionality' are not available in any related literature. This posed one of the two major motivations for this thesis; the other motivation will become clear in the course of this chapter.

The theory of multi-dimensional approximation using wavelets is developed in chapter 2. In particular, we extend the sufficient conditions given by Daubechies for 1-D wavelet frames to the multi-dimensional case in two different ways (i.e., using single and multiple dilation parameters) and show that *frames* can be constructed from a single mother wavelet. In the first case we can construct radial wavelet frames, and in the second case, the tensor product construction leads to valid frames.

The fact that wavelets are local functions suggests that learning in wavelet networks will face many of the dilemmas faced with local learning networks such as the RBFNs, local polynomial fitting, etc. A major problem in all these cases is of course the 'curse of dimensionality', which is well known across disciplinary boundaries: if one were to adhere strictly to the mathematical theory, the number of network 'units' needed becomes so excessive as to render the theory meaningless in practice. This realization has led even mathematicians schooled in rigorous theory to search for useful approximate or heuristic methods to solve complex real-world problems (see, for instance, Friedman [13] and the discussion that followed it).

In chapter 3, we use the theoretical results in chapter 2 to develop wavelet-based networks, and then examine suitable heuristic procedures to make the proposed methods practically more relevant.

Chapter 4 addresses the connection of the methods proposed in previous chapters to existing methods in diverse literature. Given the diversity and generality of neural network methods, it is not surprising that several statistical methods that have existed independently in statistics have been brought into 'neural' framework in recent years. In that spirit, we discuss the connection of the wavelet methodology to existing neural network methods and other methods such as Projection Pursuit Regression [11, 12], and Regularization Theory[27]. In particular, by simple modifications, we show that *symmetric* wavelets can be derived as special classes of the regularization network functionals.

Neural networks are in general applicable in a broad range of areas that include computer vision/image processing, adaptive signal processing (filtering,

prediction) and control. Many of the theoretical ideas that originated from control theorists have found their way into adaptive signal processing. A detailed view of the two areas can be obtained from Widrow and Stearns [33], and Astrom and Wittenmark [1]. Neural networks in *control problems* pose more challenges because of the need to prove stability, error convergence, etc. in a rigorous fashion. Moreover, neural networks can strengthen the existing connection between adaptive control and adaptive signal processing and bring them closer. This provided a parallel motivation for this thesis. The utility of the proposed methods are therefore investigated in adaptive control problems.

The literature on adaptive control and signal processing abound with techniques derived from linear systems or local linearization. What makes neural-networks attractive is their ability to solve non-linear problems in signal processing and control. As a result, in recent years, adaptive control strategies using neural-networks have been investigated by a number of researchers: Narendra and Parthasarathy [19, 20], Chen and Khalil [3], Sanner and Slotine [30], to name a few. These researchers have looked at either Multi-layer feed-forward networks or GRBFNs. Sanner and Slotine [30] also recognize the possibility of using wavelets instead of GRBFs, but to our knowledge a rigorous theory for using wavelets in a multi-dimensional network has not been developed in any work that pre-dates this work. In chapter 5, we formulate adaptive control problems and show how the wavelet network can be used in such situations.

Chapter 6 deals with results from simulation studies, the conclusions drawn, and future directions.

Chapter 2

Multidimensional Wavelets and Function Approximation

2.1 Wavelet Transform

Suppose $\psi \in L^2(\mathbf{R})$ satisfies the following admissibility condition:

$$\int_{\mathbf{R}} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty.$$

Then dilations and translations of the function ψ can be used to capture the 'local character in space and frequency' of an arbitrary function $f \in L^2(\mathbf{R})$. This property is captured in the following relations of the continuous wavelet transform (for a detailed analysis of this theory the reader is referred to [8]) .

$$f = C_\psi^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{dadb}{a^2} < f, \psi^{a,b} > \psi^{a,b}$$

where $\psi^{a,b}(x) = |a|^{-2}\psi(\frac{x-b}{a})$, and C_ψ is given by

$$C_\psi = 2\pi \int_{-\infty}^{\infty} \frac{d\omega |\hat{\psi}(\omega)|^2}{|\omega|}.$$

When one goes from the continuous case to the discrete case, the notion of *frames* becomes necessary. A formal definition of frames is given below.

Definition

Given a Hilbert Space $\mathcal{H}$, and a sequence of vectors $\{h_n\}_{n\in\mathcal{Z}}, \subset \mathcal{H}$, $\{h_n\}$ is said to constitute a frame for $\mathcal{H}$ if $\exists$ two constants $A > 0$ and $B < \infty$, such that $\forall f \in \mathcal{H}$, the following inequalities hold:

$$A\|f\|^2 \leq \sum_n | < h_n, f > |^2 \leq B\|f\|^2.$$

When the frame condition is satisfied, one can define a frame operator S as

$$Sf = \sum_n < f, h_n > h_n$$

and decompose the function f as

$$f = \sum_n w_n h_n$$

where $w_n =< f, S^{-1}h_n >$. The Hilbert space we consider is $L_2(\mathbf{R}^n)$, the space of square-integrable functions over $\mathbf{R}^n$. In the following we will consider *wavelet frames*. Daubechies has given sufficient conditions for 1-D wavelet frames. Here we extend these conditions to the multi-dimensional case in the case of joint dilations in all dimensions as well as separate dilations and translations in each dimension.

2.2 Single-scaling wavelet frame

In this section we show that it is possible to build single-scaling multi-dimensional wavelet frames by using a single mother wavelet. For this purpose we generalize

Daubechies' theorem on sufficient conditions of wavelet frame [7] to the single-scaling multi-dimensional case.

Theorem 1 *Let $\psi \in L_2(\mathbf{R}^n)$. Consider a family of dilated and translated functions of the form*

$$\Psi(a,b) = \{\psi_{l,k}(x) = a^{-\frac{1}{2}nl}\psi(a^{-l}x - bk) : \ l \in \mathbf{Z}, k \in \mathbf{Z}^n\} \tag{2.1}$$

where $x \in \mathbf{R}^n$, $a, b \in \mathbf{R}$ and $a > 1$. If the following three conditions (2.2), (2.3) and (2.4) are satisfied

$$m(\psi, a) \triangleq \operatorname*{ess\,inf}_{\|\omega\| \in [1,a]} \sum_{l \in \mathrm{Z}} |\widehat{\psi}(a^l\omega)|^2 > 0 \tag{2.2}$$

$$M(\psi, a) \triangleq \operatorname*{ess\,sup}_{\|\omega\| \in [1,a]} \sum_{l \in \mathrm{Z}} |\widehat{\psi}(a^l\omega)|^2 < \infty \tag{2.3}$$

$$\sup_{\eta \in \mathrm{R}^n} \left[(1 + \eta^T\eta)^{n(1+\epsilon)/2}\beta(\eta)\right] = C_\epsilon < \infty \text{ for some } \epsilon > 0 \tag{2.4}$$

where

$$\beta(\eta) \triangleq \sup_{\|\omega\| \in [1,a]} \sum_{l \in \mathrm{Z}} |\widehat{\psi}(a^l\omega)| \cdot |\widehat{\psi}(a^l\omega + \eta)| \tag{2.5}$$

then there exists $b_0 > 0$ such that $\forall b \in (0, b_0)$, the family $\Psi(a,b)$ in (2.1) constitutes a frame of $L^2(\mathbf{R}^n)$, in other words, there exist two constants $A > 0$ and $B < +\infty$, such that $\forall f \in L^2(\mathbf{R}^n)$, the following inequalities hold

$$A\|f\|^2 \leq \sum_{l,k} |\langle \psi_{l,k}, f \rangle|^2 \leq B\|f\|^2$$

where the sum ranges are $l \in \mathbf{Z}$ and $k \in \mathbf{Z}^n$, $\langle \cdot, \cdot \rangle$ denotes the inner product in $L_2(\mathbf{R}^n)$. □

Note that for the family $\Psi(a,b)$ defined by (2.1), the dilation index l is a scalar, and the scalar dilation parameter a^l is shared by all the dimensions of a wavelet. The proof of this theorem is given in Appendix A.1.

2.3 Multi-scaling wavelet frame

We introduce the dilation and translation matrices D_j and T as

$$D_j = diag\left(a^{j_1}, \cdots, a^{j_n}\right)$$

where

$$j = (j_1, \cdots, j_n)^T \in \mathbf{Z}^n$$

and

$$T = diag\left(b_1, \cdots, b_n\right).$$

With D_j and T thus defined, separate dilation and translation parameters can be used in wavelet functions. The following theorem is an analog of Theorem 1 in the multi-scaling case.

Theorem 2 *Let $\psi \in L_2(\mathbf{R}^n)$. For $a \in \mathbf{R}$, $a > 1$, $b = (b_1, \cdots, b_n) \in \mathbf{R}^n$, and $b_i > 0$, $i = 1, \cdots, n$, consider the family of translated and dilated functions of the form*

$$\Psi(a, b) = \{\psi_{j,k}(x) = \det D_j^{\frac{1}{2}} \psi(D_j x - Tk) : j, k \in \mathbf{Z}^n\}.$$

If

$$m(\psi, a) \triangleq \operatorname*{ess\,inf}_{|\omega_i| \in [1,a], i=1,\cdots,n} \sum_{j \in Z^n} |\widehat{\psi}(D_{-j} w)|^2 > 0,$$

$$M(\psi, a) \triangleq \operatorname*{ess\,sup}_{|\omega_i| \in [1,a], i=1,\cdots,n} \sum_{j \in Z^n} |\widehat{\psi}(D_{-j} w)|^2 < \infty$$

and

$$\sup_{\eta \in \mathrm{R}^n} [(1 + \eta^T \eta)^{n(1+\epsilon)/2} \beta(\eta)] = C_\epsilon < \infty \text{ for some } \epsilon > 0$$

where

$$\beta(\eta) \triangleq \sup_{|\omega_i| \in [1,a], i=1,\cdots,n} \sum_{j \in \mathbf{Z}^n} |\widehat{\psi}(D_{-j}w)|.|\widehat{\psi}(D_{-j}w+\eta)|,$$

then there exists[1] $b_0 > 0$ *such that* $\forall b \in (0, b_0)$, *the family defined above constitutes a frame for* $L_2(\mathbf{R}^n)$; *i.e.* , $\exists$ *two constants* $A > 0$ *and* $B < \infty$, *such that* $\forall f \in L_2(\mathbf{R}^n)$, *the following inequalities hold*

$$A\|f\|^2 \leq \sum_{j,k} |\langle \psi_{j,k}, f \rangle|^2 \leq B\|f\|^2$$

□

The proof of this theorem is given in Appendix A.2.

2.4 Construction of wavelet frames

We are interested in a methodology that allows us to construct the multi-dimensional wavelet function leading to frames; i.e., the problem is to find a wavelet function that satisfies, together with its dilation and translation parameters, the sufficiency conditions outlined in the above theorems. In this section we first consider the tensor product construction of multi-scaling wavelet frames; then we discuss possible non product constructions.

2.4.1 Tensor product frames

Let $\psi(x)$ be a tensor product of 1-dimensional wavelet functions, i.e.,

$$\psi(x) = \psi_1(x_1) \cdots \psi_n(x_n)\,.$$

[1] abusing notation, we consider element-wise bounds when we refer to vector bounds in this thesis.

Then,

$$\widehat{\psi}(w) = \widehat{\psi}_1(\omega_1) \cdots \widehat{\psi}_n(\omega_n)\,.$$

$\psi_i(x_i), i = 1, \cdots, n$, must satisfy the admissibility condition:

$$\int \frac{|\widehat{\psi_i}(\omega_i)|^2 \mathrm{d}\omega_i}{|w_i|} < \infty.$$

Under mild conditions of decay, this is satisfied if we choose $\psi_i(x_i)$ such that

$$\int \psi_i(x_i) \mathrm{d}x_i = 0.$$

If these 1-dimensional functions can constitute frames, they must satisfy the first two conditions outlined in theorem 2, as applied to the 1-dimensional case, which are necessary conditions as well [8].

Moreover, Daubechies [8] shows that in 1-D, a single sufficient condition on the decay of ψ_i as given by

$$|\widehat{\psi_i}(\omega_i)| \leq C_i |\omega_i|^{\alpha} (1 + |\omega_i|^2)^{-\frac{\gamma}{2}} \text{ with } \alpha > 0 \text{ and } \gamma > \alpha + 1$$

is equivalent to the second and third conditions of the theorem.

Since in practice this decay condition is rather mild, for the purpose of construction, we assume that it is satisfied and hence all conditions of the theorem are satisfied.

Hence in the multidimensional case, by using the inequalities in 1-D above, and the fact that the infimum and supremum can now be taken over the sum in each dimension, we have

$$
\begin{aligned}
m(\psi, a) &= \operatorname*{ess\,inf}_{|\omega_i| \in [1,a], i=1,\cdots,n} \Big\{ \sum_{j_1} |\widehat{\psi}_1(a^{-j_1}\omega_1)|^2 \cdots \sum_{j_n} |\widehat{\psi}_n(a^{-j_n}\omega_n)|^2 \Big\} \\
&> 0,
\end{aligned}
$$

and

$$
\begin{aligned}
M(\psi, a) &= \operatorname*{ess\,sup}_{|\omega_i| \in [1,a], i=1,\cdots,n} \Big\{ \sum_{j_1} |\widehat{\psi}_1(a^{-j_1}\omega_1)|^2 \cdots \sum_{j_n} |\widehat{\psi}_n(a^{-j_n}\omega_n)|^2 \Big\} \\
&< \infty.
\end{aligned}
$$

For the third condition, we have the following inequality,

$$
\begin{aligned}
\sum_{|k| \neq 0} \Big[\beta(2\pi T^{-1}k) \beta(-2\pi T^{-1}k) \Big]^{1/2} &\leq \sum_{k_1} (1 + (2\pi b_1^{-1} k_1)^2)^{-\frac{(1+\epsilon)}{2}} \cdots \\
&\quad \sum_{k_n} (1 + (2\pi b_n^{-1} k_n)^2)^{-\frac{(1+\epsilon)}{2}} \\
&< \sum_{k_1} |2\pi b_1^{-1} k_1|^{-(1+\epsilon)} \cdots \sum_{k_n} |2\pi b_n^{-1} k_n|^{-(1+\epsilon)}.
\end{aligned}
$$

Since each sum over k_i converges, $i = 1, \ldots, n$, we have that the sum involving β converges. Moreover, as $b_i \to 0$, $i = 1, \cdots, n$, this sum tends to 0. Hence all conditions of the theorem are satisfied.

Therefore the tensor product construction leads to valid frames of wavelets.

2.4.2 Necessary conditions

To make the results complete, we are interested in obtaining necessary conditions as in the 1-D case. In particular, it would be appropriate to check whether the admissibility condition for discrete wavelet frames has the same structure as

continuous wavelets in the multidimensional case. In the tensor product set up, this follows trivially since the 1-D admissibility conditions lead to

$$\int_{\mathrm{R}^n} \psi(x)\mathrm{d}x = 0.$$

From recent extensions on the bounds for 1-D case [5, 7], the following holds for any frame $\psi_{i_{j_i,k_i}}$ ($i = 1, \cdots, n$ is the dimension index, j_i, k_i are dilation and translation indexes respectively):

$$A_i \leq \frac{2\pi}{b_i} \sum_{j_i} |\widehat{\psi_i}(a^{-j_i}w)|^2 \leq B_i\,.$$

Considering multiplication of the above inequalities over $i = 1, \cdots, n$, we have

$$A = A_1 \cdots A_n \leq (2\pi)^n \det T^{-1} \sum_{j} |\widehat{\psi}(D_{-j}w)|^2 \leq B_1 \cdots B_n = B$$

In [5], this bound is obtained for the case of *Riesz bases*. However, the proof relies only on the frame condition, and therefore the above inequality is general in that it holds for arbitrary frames (not necessarily of the tensor product type). Another problem is to construct such arbitrary frames.

2.4.3 Non-separable frames

The observation that all conditions of the theorems on sufficient conditions hinge on the boundedness and decay of terms involving $|\widehat{\psi}(\cdot)|$ suggests the possibility of multiplying the tensor product wavelet in the frequency domain by a function of the form

$$p(w) = \sum_{l \in \mathrm{Z}^n} c_l\, e^{-il^T w},\ c_l \in \mathbf{R}$$

which can be the Fourier series of a periodic function.

Let the new wavelet be

$$\widehat{\psi}_p(\omega) = p(w)\widehat{\psi}(\omega)$$

where $\psi(\cdot)$ corresponds to the wavelet constructed as a tensor product. If

$$0 < \sum_l |c_l| < \infty\,,$$

then the fact that

$$0 < |\widehat{\psi}_p(\omega)| \leq (\sum_l |c_l|)|\widehat{\psi}(\omega)|$$

implies that all conditions of Theorem 2 are satisfied.

Therefore, one can construct non-tensor product wavelet frames from the tensor product frames. In the case of Riesz bases, similar results are obtained in [5].

The 1-D wavelet function could be the Mexican hat, a combination of a few sigmoids (e.g.[23]) etc. The choice of the wavelet used in networks for learning is dictated by considerations of smoothness, implementability in analog hardware, separability, etc. Some of these issues will be considered in chapter 3.

Radial Wavelet Frames

When we impose radial symmetry on the mother wavelet, $\hat{\phi}(\omega) = \hat{\phi}(||\omega||)$ the following isotropic admissibility condition results,

$$\int_0^\infty \frac{dh}{h}|\widehat{\phi(h\omega)}|^2 < \infty.$$

For instance, the radial Mexican hat function $\phi(x) = (n - ||x||^2)e^{\frac{-x^2}{2}}$ is used by many researchers when continuous wavelet transforms are considered, and in

particular the difference of Gaussians as approximation to the Laplacian of the Gaussian is popular in computer vision applications. In the *discrete* case, such a radial construction is implied in the conditions of Theorem 1. It is easy to see that the first two conditions follow directly. Moreover, if the 1-dimensional mother wavelet is chosen according to the mild decay conditions (Daubechies [8]), i.e.,

$$|\widehat{\psi_i}(\omega_i)| \leq C_i|\omega_i|^{\alpha}(1+|\omega_i|^2)^{-\frac{\gamma}{2}} \text{ with } \alpha > 0 \text{ and } \gamma > \alpha + 1$$

then the third condition of Theorem 1 is also satisfied. Therefore, the construction involves the following:

1. Select a symmetric 1-D mother wavelet $\phi(x)$ and calculate the Fourier Transform $\hat{\phi}(\omega)$.

2. Let the multi-dimensional wavelet satisfy

$$\hat{\phi}(\omega) = \hat{\phi}(\|\omega\|).$$

   The Inverse Fourier Transform of $\hat{\phi(\omega)}$ gives the radial mother wavelet candidate for higher dimensions.

In the sequel, we shall be concerned with tensor-product wavelets. Once a frame is selected, for any $f \in L^2(\mathbf{R}^n)$ we can write,

$$f = \sum_{m,n} c_{mn}\psi_{mn}$$

where

$$\psi_{mn}(x) = detD^{1/2}\psi(Dx - Tb),$$

$$D = diag(a[1]^m, \cdots, a[N]^m)$$

and $T = diag(n[1], \cdots, n[N])$. The coefficients c_{mn} represent local information at the space-frequency location of m, n. Therefore it is desirable to define the centers of time-frequency in a rigorous fashion.

The following definitions for the n-dimensional case are appropriate.

$$x_c(f) = \frac{1}{\|f\|^2} \int_{\mathbf{R}^n} x_1 \cdots x_n |f(x)| dx$$

$$\omega_c(|\hat{f}|) = \frac{1}{\|\hat{f}\|^2} \int_{[0,\infty)^n} \omega_1 \cdots \omega_n |\hat{f}(\omega)|^2 d\omega$$

For the tensor product wavelet, the center co-ordinates are the centers in each dimension.

In practice, functions are essentially concentrated in a spatio-spectral region in the following sense.

$$\int_{x_l \le x \le x_u} |f(x)|^2 dx \ge (1 - \epsilon) \|f\|^2 \tag{2.6}$$

$$\int_{\omega_l \le |\omega| \le \omega_u} |\hat{f}(\omega)|^2 d\omega \ge (1 - \epsilon) \|f\|^2 \tag{2.7}$$

Hence the set of (m, n) is truncated to a finite index set $\mathcal{I}$ and we need to be precise about the truncation error in truncating with finite (m, n) . We work with the tensor product construction. We note again the abuse of notation in using inequalities and bounds in the vector case, i.e., since $x \in \mathbf{R}^N, \omega \in \mathbf{R}^N$, the vector inequalties involving x and ω are taken elementwise. First we note that $c_{mn} =< f, S^{-1}\psi_{mn} >$, where S is the operator in connection with frames defined earlier. $S^{-1}\psi_{mn}$ is called the dual frame $\tilde{\psi}_{mn}$. Hence

$$\|f - \sum_{m,n \in \mathcal{I}} < f, \tilde{\psi}_{mn} > \psi_{mn}\| = \sup_{\|h\|=1} | < f, h > - \sum_{m,n \in \mathcal{I}} < f, \tilde{\psi}_{mn} >< \psi_{mn}, h > |$$

$$
\begin{aligned}
&= \sup_{\|h\|=1} \Big| \sum_{m,n \notin \mathcal{I}} < f, \tilde{\psi}_{mn} >< \psi_{mn}, h > \Big| \\
&\leq \sup_{\|h\|=1} \sum_{m<m_l, m>m_u} \sum_{n \in Z^n} (\cdot) \\
&\quad + \sup_{\|h\|=1} \sum_{m_l \leq m \leq m_u} \sum_{a^{-m}x_u + x_h^+ < nb < a^{-m}x_l - x_h^-} (\cdot)
\end{aligned}
$$

Here x_h^+, x_h^- are used to consider a small region just beyond the boundaries in spatial region defined by the set $\mathcal{I}$.

Define a cover B_ϵ to $\mathcal{I}$ as

$$
B_\epsilon(\omega_l, \omega_u; x_l, x_u) = \begin{cases} (m,n) \in \mathrm{Z}^n; & m_l \leq m \leq m_u, \text{and} \\ a^m x_l - x_h^-(\epsilon, m_l, m_u) < nb & \leq a^{-m}x_u + x_h^+(\epsilon, m_l, m_u). \end{cases}
$$

The notion here is that there exist lower bounds on x_h^+, x_h^-, m_u and upper bound on m_l to meet the essential spatio-spectral concentration of the function to the region $[x_l, x_u] \times [w_l, w_u]$. Using the Cauchy-Schwarz inequality, we can derive the following, paralleling the one-dimensional case studied in [8]. Coarse estimates for $m_{ui}, m_{li}, x_{hi}^+, x_{li}^-$ that depend on the decay factor of the mother wavelet and the spetral limits are used in this derivation to show the desired results.

$$
\|f - \sum_{m,n \in B_\epsilon(\omega_l, \omega_u; x_l, x_u)} < f, \tilde{\psi}_{mn} > \psi_{mn}\| \leq \sqrt{B/A} \left[\begin{array}{l} \left(\int_{\omega \notin [\omega_l, \omega_u]} d\omega |\hat{f}(\omega)|^2 \right)^{\frac{1}{2}} \\ + \left(\int_{x \notin [x_l, x_u]} dx |f(x)|^2 \right)^{\frac{1}{2}} \\ + \epsilon \|f\| \end{array} \right]
$$

If we use the definitions in 2.6, 2.7, with the same ϵ, we have that

$$
\|f - \sum_{m,n \in B_\epsilon} c_{mn} \psi_{mn}\| = O(\epsilon),
$$

which is the desired result consistent with the intuition of essential spatio-spectral concentration, i.e., it suffices to use the nodes that fall close to the spatio-spectral region.

The number of nodes required is given by

$$\begin{aligned}
\#B_\epsilon &= \prod_{i=1}^{N} \sum_{m_i=m_{il}}^{m_i=m_{iu}} \left|\frac{a^{m_i} x_{ui} + x_{hi}^{+}}{b}\right| + \left|\frac{a^{m_i} x_{li} - x_{hi}^{-}}{b}\right| \\
&= \prod_{i=1}^{N} b^{-1} \left(x_{iu} - x_{il}\right) a^{m_{li}} \left(\frac{a^{m_{ui}-m_{li}} - 1}{a-1}\right) \\
&\qquad \text{neglecting effects of } x_{hi}^{+}, x_{hi}^{-} \ .
\end{aligned}$$

This detailed theory provides the means for *uniformly* approximating any function $f \in L^2(\mathbf{R}^N)$ to a desired degree of accuracy, and translates into a network formulation. Such a formulation is studied in the following chapter.

Chapter 3

Wavelet-based Networks and Learning

The methodology for learning involves training a network either on-line or off-line based on a set of observations, so that the resulting error in approximation is within acceptable limits. Such networks should be able to 'generalize', meaning that when presented with input not used in training, the network should be able to approximate the mapping well. This ability is the key to learning, but achieving this is the most difficult part of the learning process since the training set used cannot in general adequately represent the whole input space. We will formalize these notions and consider the issues arising out of this dilemma.

3.1 Generalization Error, Network Structure and the Size of the Training Set

The generalization error is defined as

$$\|f - f_{net}\| = \epsilon_{gen}$$

where f is the desired mapping $\mathbf{R}^n \mapsto \mathbf{R}$, and f_{net} is the mapping provided by the network. Now, we note that the learning process gives rise to two distinct

types of errors, viz.

1. The *approximation error*, $f - f_{approx}$, which results from the fact that a finite amount of resources (nodes or neurons) are used to approximate the function. This tells us that approximation theory can be used as a means of determining the size of the network to be used.

2. The *estimation error*, $f_{approx} - f_{net}$, which results from the fact that the co-efficients or weights are *estimated* from a finite amount of data. Thus the nature and size of the training set used has to come from statistical considerations.

Thus one can write,

$$\epsilon_{gen} \leq \|f - f_{approx}\| + \|f_{approx} - f_{net}\| \tag{3.1}$$

$$= \epsilon_{approx} + \epsilon_{est} \tag{3.2}$$

However, in this thesis we will not obtain statistical bounds on these errors. Recent papers by Barron [2], Niyogi and Girosi [22] provide good analyses in similar contexts in neural networks, and it should be possible to perform similar analyses in our case. It suffices to note that motivations for these analyses come from fundamental problems in "learning", viz., how to trade off the above two errors, i.e., ϵ_{approx} and ϵ_{est} so that the resulting error ϵ_{gen} is within acceptable limits. These questions are related to determining the network complexity (i.e., network size) and the size of the training set (sample complexity) that are optimal. Results reported in the literature agree with the empirical evidence that several combinations of these two parameters optimally result in the same generalization error. As a result, the choices are in practice determined by the

availability of resources– for network units and for data collection. In other words, if a large amount of data can be obtained, it is possible to use this high "information content" or "feature content" to obtain more compact networks. Conversely, when only a small amount of data is available, it is possible to arrive at the same generalization error by using a larger network. Again, this assertion is supported by theoretical results as well as empirical observations (simulations).

This discussion suggests one among several important reasons to develop *on-line sequential learning* strategies that build near-optimal networks provided sufficiently large amount of data. Because the data are presented only once [1], there is no need to store a large amount of data. Other reasons will become apparent in later sections of this chapter.

3.2 Local versus Global Learning

Learning schemes based on 'global' and 'local' learning schemes have distinct properties, most of which are well known in the literature. In global learning, no association can be made between a subset of the input space and the adjustable elements (weights). At every instant of weight adjustment, all weights get adjusted. This has the advantage of resulting in a compact network and better generalization, but one has to contend with poor accuracy and sensitivity. The multi-layer sigmoidal networks in widespread use are global learning networks. In contrast, local learning is characterized by weights corresponding to a small region of the input space, a higher degree of accuracy, a smaller number

[1] repeated presentation may be necessary when the same strategies are used off-line, but a small amount of data is sufficient in this case

of weight adjustments, etc on the positive side and a larger number of units, poor generalization capability because of too close fitting, etc, on the negative side. However, there are certain applications, where a local mapping is highly desirable. When training data are obtained from on-line interactions with the system to be modelled, the training samples may tend to fixate to a certain region of the input space. This can harm the generalization capability in global learning since all weights are repeatedly adjusted (see, for instance, [10] for a discussion of some related issues in learning control applications). In certain cases, accuracy is the predominant concern and the requirement for larger memory is acceptable. In other cases where local learning is essential, but large memory cannot be accommodated, techniques for reducing the 'curse of dimensionality' need to be developed. The Gaussain Radial Basis Function networks, and the Wavelet Networks (WNs) used in this work are local learning networks.

3.2.1 Training Local Learning Networks

A claim is often made that the linear-in-the parameter structure of local learning networks (such as the GRBFNs and WNs) simplifies training compared to the extremely slow *back-propagation* procedure used in sigmoidal neural-networks. Such simplicity is deceptive unless techniques that address the issue of how to select the 'basis' or 'receptive field' functions are developed. In many problems, this can again necessitate gradient based nonlinear optimization procedures. In the wavelet frame work, this problem comes down to selecting the appropriate sets of dilations and translations (also referred to as the dictionary) in an efficient manner. Once this is done, training is of course easier than back-propagation,

and several methods can be used for determining the coefficients depending on whether the training is on-line or off-line and the amount of computations. In the next few sections of this chapter we give a details of our methodology.

3.2.2 Theoretical Difficulties of Network Construction

Existing methodologies to construct sigmoidal networks are lacking in several respects: to rephrase the discussion at the start of the chapter, for a specified tolerance, questions such as how many nodes are necessary in a hidden layer?, how many layers are necessary?, and how many training samples are needed?, have only partial answers at the present time. Several researchers have recently shown that feed-forward neural networks are universal approximators (see, e.g.,[16]), and that a single hidden layer is sufficient to approximate any arbitrary nonlinear function to a desired accuracy provided a sufficient number of neurons are used; however sufficiency does no lead to any rigorous procedures for network construction.

Why use wavelets ?

Pati and Krishnaprasad [23] circumvented these questions to some extent by using discrete wavelet transforms in place of sigmoids; however, their work is primarily applicable to 1-D problems; adequate treatment of multi-dimensional wavelet theory, and problems faced due to the 'curse of dimensionality' are not available.

One advantage with wavelets is that the problems of input pre-processing, scaling etc, are avoided (inherently structured in this way). Similarity of

wavelets to Radial Gaussians also makes wavelets attractive in applications where Gaussians have been successfully employed. Indeed, wavelet theory provides a natural basis for multi-resolution hierarchical schemes, unlike the artificially imposed multi-resolution hierarchies in the case of Gaussians as used in vision applications. Such a multi-resolution scheme is intuitively appealing since they possess the ability to zoom-in on areas of high-frequency concentration, analogous to the way the human brain processes information. Moreover, because of the above structure, wavelets offer the possibility of obtaining more compact networks though this is not a universally applicable claim.

3.2.3 Problems Faced in High Dimensions

Many wavelet/neural learning problems often attacked in the literature concern one-dimensional applications, with no straightforward extension to high dimensional cases. Such studies have very little utility in practical problems since neural networks find their usefulness predominantly in high-dimensional applications. High-dimensional problems pose more challenges and much remains to be done in the direction of achieving good generalization at significantly reduced complexity. If on-line sequential adaptation is attempted in à high dimensional setting, the problems become still more complicated. In particular, traditional methods would require storage of large amounts of past data, which is difficult, and moreover the function to be learnt can be highly non-stationary. Information theoretic considerations such as Cross-Validation/Generalized Cross Validation, Minimum Description Length Principle or Akaike's Information Criteria are not applicable sequentially on-line because of the repetitive nature of these nonlinear

optimization procedures.

3.2.4 Training Local Learning Networks

A claim is often made that the linear-in-the parameter structure of local learning networks (such as the GRBFNs and Wavelet Networks) simplifies training compared to the extremely slow back-propagation procedure used in sigmoidal neural networks. Such simplicity is deceptive unless techniques that address the issue of how to select the 'basis' or 'receptive field' functions are developed. In many problems, this can again necessitate gradient based non-linear optimization procedures. In the wavelet frame work this problem comes down to selecting the appropriate sets of dilations and translations (also referred to as the dictionary) in an efficient manner. Once this is done, several methods can be used for determining the coefficients depending on whether the training is on-line or off-line and the amount of computations. In the next few sections of this chapter we give details of our methodology.

We have shown in chapter 2, how to construct multi-dimensional wavelet frames, and how these frames can be used to approximate functions in high-dimensional spaces. This theory maps directly into a network configuration.

3.2.5 Determining the Spatio-Spectral Centers and the Coefficients

The first issue in the wavelet network construction is to determine the frequency content of the system function, which is assumed to be essentially band-limited

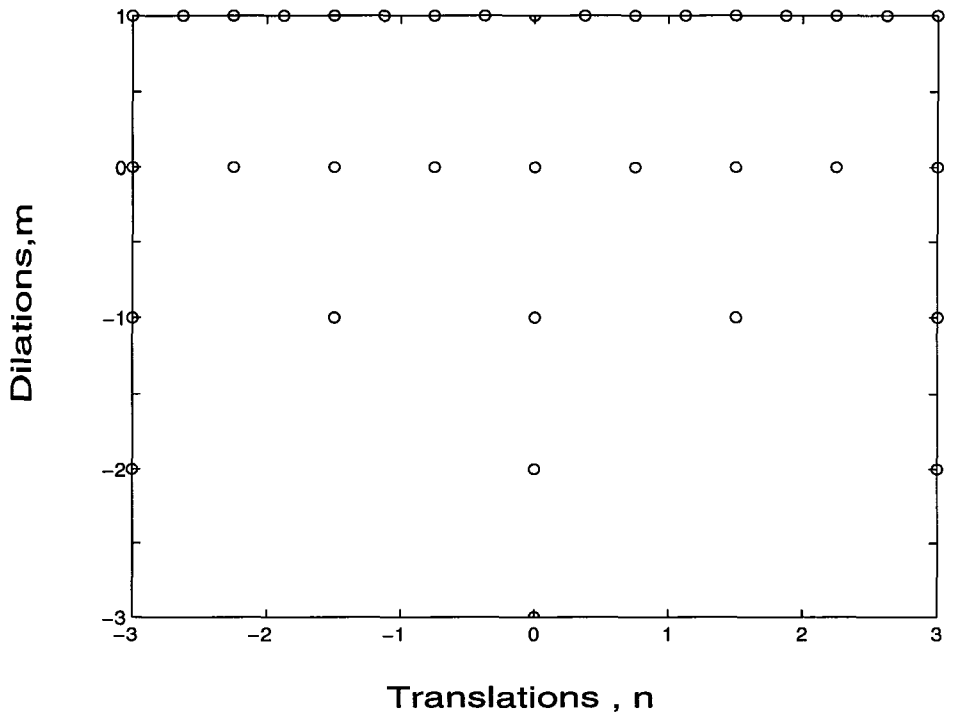


Figure 3.1: Spatio-Spectral Distribution

in space and frequency in $[\omega_l, \omega_m] \times [x_l, x_m]$. If somehow this information is known *a priori* the space-frequenecy region on which the approximation is to be attempted becomes clear. Theoretically, one would then try to select all the spatio-spectral centers that fall within the region of interest. The computation of coefficients would be straightforward theoretically. However, the number of such centers increases enormously with each additional dimension, and this would make the methodology devoid of any practical merits. Several approaches can be followed to give realistic solutions depending on the nature of the problem in hand:

1. The dimension of the input space

2. The amount of data and whether on-line or off-line

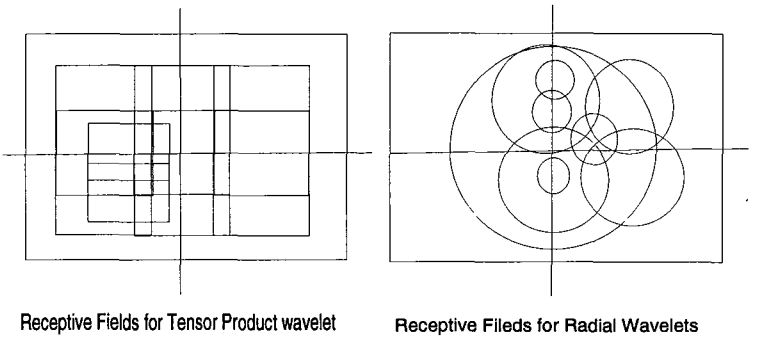


Figure 3.2: Typical Receptive Fields in Two Dimensions

3. The degree of approximation desirable

4. *A priori* knowledge of smoothness information if any.

In practice attempts at calculating the spectral content based on the training data are bound to be fruitless unless one considers low dimensions and a small set of training data. The methodology we propose in the next section obviates the need to perform such calculations of spectral content. We assume that there is no significant constraint on the amount of data that can be gathered on-line.

3.3 A Heuristic Methodology for Dynamic Selection

Consider each dimension separately; the same procedure will be performed in each dimension. Let $L(l), M(l), l = 1, \cdots, n$ be the low and high frequency limits. Choose $L(l)$ corresponding to the case in which entire spatial region $[x_l(l), x_m(l)]$ is covered by two nodes (translations), i.e.,

$L(l) = -\frac{\log(x_m(l) - x_l(l)) - log(b(l))}{log(a)}$.

There is no need to know $M(l)$ except that knowledge of $M(l)$ can provide a rough upper bound on the number of dilation levels to be used. But this is not essential to the method.

It is now possible to build on successive levels of dilations on-line. We observe that in figure 3.1, the width between adjacent nodes for a dilation level m is given by $a^{-m}b$. This information can be used with information on the nearest neighbour nodes, to develop the following strategy. Theoretical justification will be given in the next section.

- Initialize:

$$m(l) = L(l)$$
$$d(l) = \frac{a^{-m(l)}b(l)}{2}$$

  Set first node (translation value) to $\frac{x(l)a^{L(l)}}{b(l)}$ rounded to the nearest integer.

- begin on-line; determine $nearest(l)$ (The distance to the nearest existing translation node).

  If $nearest(l) > d(l)$[2] and $|ynet - y| > \epsilon$ add a new translation node at

$$n(l) = \frac{x(l)a^{L(l)}}{b(l)} \text{ rounded.}$$

- Select the current weight for the new node as $c_{current} = \frac{|ynet-y|}{p(x)}$ where $p(x) = \psi_1 \cdots \psi_n$, calculated at spatio-spectral locations $m(l), n(l)$.

If a new node is not selected for the present data, adapt the coefficients using the LMS[3] [32] algorithm.

When the network has learned sufficiently at this level (this is determined by

[2]Considerations on this choice are detailed in the next section

[3]Other algorithms such as RLS-Kalman or variants thereof can also be used at the expense of increased complexity of implementation

the fact that no new node is added for a sufficient number of continuous data points, as shown by a flag) set $m(l) = m(l) + 1$;
Continue this on-line. After a sufficient number of nodes are learnt, the algorithm automatically stops adding new units.

If training is desired off-line, optimization of this initial model can be performed based on orthogonal least squares (OLS) [4] or the orthogonal matching pursuit (OMP) [24], which is similar to OLS.

3.4 Justification

The methodology proposed above can be justified based on geometric model growth. A work in similar spirit for Platt's RAN can be found in [17].

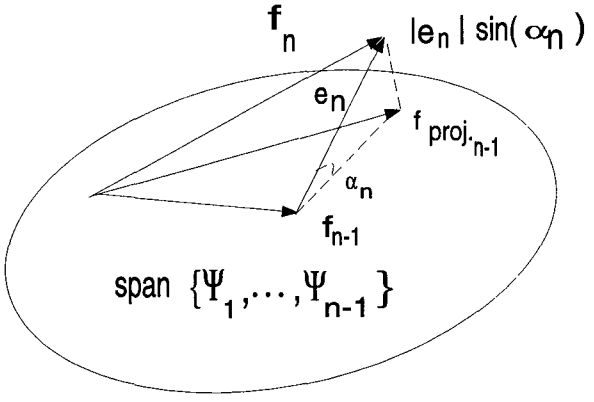


Figure 3.3: The Geometric Picture

The geometric picture is illustrated in figure 3.3. Notice that $f_n = \sum_{i=1}^{n} c_i \psi_i$ implies that

$$f_n = f_{n-1} + e_n \quad \text{with } e_n = c_n \psi_n$$

The best approximation $f_{proj_{n-1}}$ to f that one can get from a set of frames

$\lvert e_n \rvert$	$\lvert \sin(\alpha_n) \rvert$	$\lvert e_n \rvert \lvert \sin(\alpha_n) \rvert$	Decision	Interpretation
$< \epsilon_1$	$< \epsilon_2$	$< \epsilon_1\epsilon_2$	Use LMS	The Projn. is sufficient.
$< \epsilon_1$	$\geq \epsilon_2$	$< \epsilon_1$	Use LMS	The Projn. is sufficient
$\geq \epsilon_1$	$< \epsilon_2$	?	Use LMS,flag	need more data [4]
$\geq \epsilon_1$	$\geq \epsilon_2$	$\geq \epsilon_1\epsilon_2$	Add a new node	The Projn. is inadequate

Table 3.1: Possible Combinations and Actions

$\{\psi_i, i = 1, \cdots, n-1\}$ is the projection of f onto the space spanned by the set $\{\psi_i, i = 1, \cdots, n-1\}$. Our first problem is to decide whether a new 'basis unit' needs to be added at this stage. Such a decision obviously can be based on whether the projection onto the span is inadequate, i.e., whether $\|f_n - f_{proj_{n-1}}\| = |e_n| \sin(\alpha_n)$ exceeds an allowable threshold ϵ. Here $|e_n| = \|f_n - f_{n-1}\|$.

Table 3.1 shows the four possibilities. Only the fourth case warrants addition of a new 'unit', while the third case suggests via a flag that new dilation levels may be needed at the particular locality, or at least more data are needed. Therefore we can separate the condition in terms of $|e_n|$ and $|\sin(\alpha_n)|$. Now it is difficult to calculate α_n. However, notice that $e_n \cos(\alpha_n)$ lies in the span of $\{\psi_1, \cdots, \psi_{n-1}\}$. Therefore the condition on $\sin(\alpha_n)$ can be recast as

$$sup_{i=1,\cdots,n-1}\{|\frac{< \psi_n, \psi_i >}{\|\psi_n\|\|\psi_i\|}|\} \leq 1 - \delta \text{ where } \delta \leq 1 \ .$$

3.4.1 Interpreting the Condition on $< \psi_n, \psi_i >$

For this calculation we make a choice for the wavelet with the understanding

[4] In addition to finding the projection, the flag indicates that at this local region additional nodes may be required

that the same method can be used for other wavelet functions. We consider the Mexican hat $(1-x^2)\,e^{-\frac{x^2}{2}}$. The tensor-product of this wavelet in two dimensions is shown in figure 3.4.

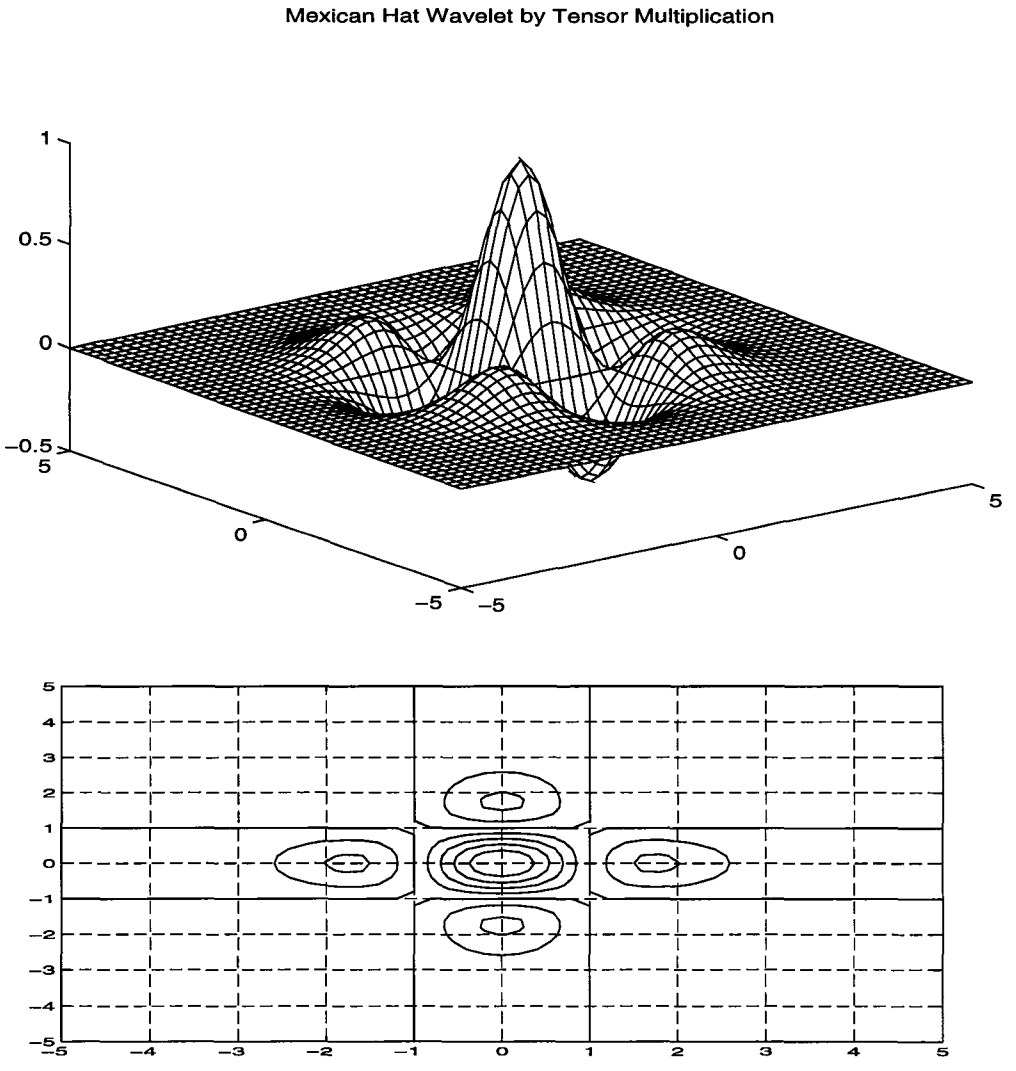


Figure 3.4: The Tensor Product of Mexican Hat

Also we note that our tensor product results in

$$< \psi_n, \psi_i > = < \psi_{n1}, \psi_{i1} > \cdots < \psi_{nN}, \psi_{iN} > .$$

Therefore by calculations shown in Appendix A.3 we arrive at the following reduction. $\forall j \in \{1, \cdots, n-1\}$,

$$< \psi_{ij}, \psi_{nj} >= e^{-\frac{a^{2m} d_j^2}{4}} \left(1.0 - a^{2m} d_j^2 + \frac{1}{12} a^{4m} d_j^4\right)$$

where

$$d_j = x_{ji} - x_{jn}.$$

It is interesting to note that graphically $< \psi_n, \psi_i >$ takes the form of the mother wavelet and that $\cos(\alpha_{nj})$ can take negative values (i.e., the angle α_n can be greater than $\pi/2$). In the case of Gaussians (GRBFNs), we can easily verify that the form of this product is again Gaussian, and that $0 \leq \alpha_n < 90°$. This shows clearly the differences between the Gaussian and the Wavelet cases. An important property emerges here: if we maintain the distance d at zero-crossing points, we get $\alpha_n = 90°$: orthogonality between nodes. However notice that this orthogonality doesn't hold across all multiples of the distance since the function has only four (symmetric) zero-crossings in each dimension that are not integer multiples and hence cannot generate a regular lattice. Also notice that between the first and second zero-crossings, the absolute value of $\cos(\alpha_n)$ can be high. Since near-orthogonality is desirable, the fore-mentioned observations suggest that one can choose the distance either to co-incide with any of the zero-crossings or to be around them. In the case where a condition of the form $|d| > d_0(m)$ where $d_0(m)$ is a fixed distance for a given m, is desired it is possible to attempt $d_0(m) \geq$ the distance to the second zero-crossing.

In the case of Gaussians, such orthogonality between any nodes obviously cannot exist. Choice of distance for near orthogonality is possible. It has been

brought to our notice that in a different context, Holcomb and Morari [15] suggested some *ad hoc* procedures in related spirit to this work for forcing orthogonality by using particular penalty terms that help spread out the basis centers in RBFN learning.

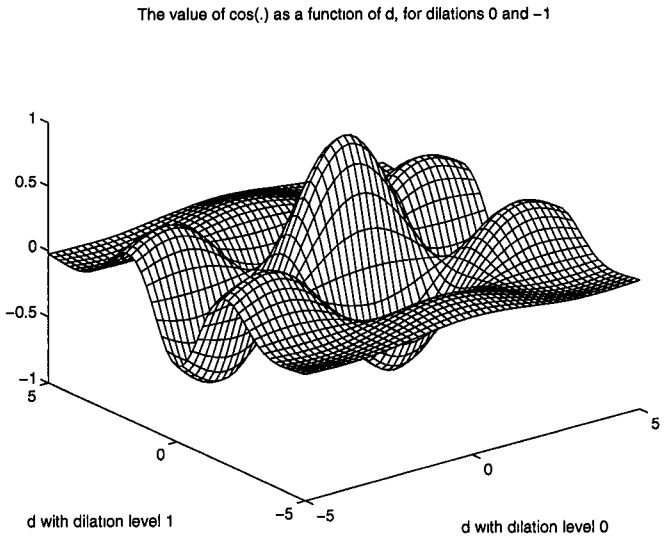


Figure 3.5: The Variation of $\cos(\alpha_n)$ with d, in $2D$

Figure 3.5 shows the value of $\cos(\alpha_n)$, which results from the multiplication of the values using separate dilations in each dimension for the 2-Dimensional case.

3.5 Implementation Issues

An important consideration in implementation is separability. Separability of the wavelet makes the network more amenable to parallel hardware implementation. Since sigmoids are more easily implemented than Gaussians or Mexican Hat, wavelet frames can be constructed by superposition of sigmoids. Pati and

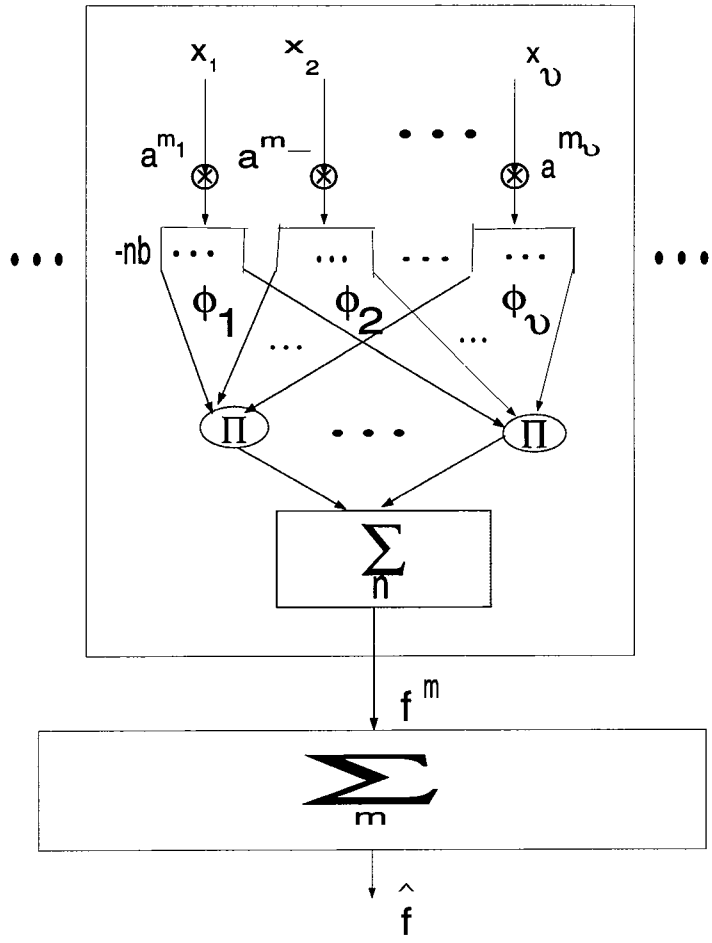


Figure 3.6: The Wavelet Network

Krishnaprasad [23] give more details on numerical procedures for constructing frames from sigmoids. Separability can also be used to advantage in constructing efficient algorithms such as the LMS Tree (see for instance [29]) under certain restrictive conditions. Since we are interested in more general cases, we choose to implement the sequential learning strategy we develop. We may note that the Gaussian RBF networks are separable and they allow the addition of successive dimensions consistent with biological observations[27]. However, as we have shown in chapter 2, Radial Wavelets are not in general known to be separable

frames.

Choice of the wavelet should be based on the expected structure of the function. Smoothness and symmetry should be considered. For logical functions that involve switching between logic states, the Haar Basis, which happens to be orthonormal as well, though not symmetric, is a sensible choice. Many other orthonormal wavelet frames have complicated expressions, and large supports. They do not lend themselves readily to hardware implementation, and large receptive fields could be detrimental to attempting local learning.

Because we assume piece-wise smooth nonlinear functions with high spatial variability, we choose to experiment with the Mexican Hat in a tensor product form. Although using single-scaling wavelets in radial form can simplify the learning and reduce the number of network nodes, we would like to test the effectiveness of our methods under more general conditions.

For the Mexican Hat wavelet, from numerical calculations given in Daubechies [8], for the dilation $a = 2$, values of $b > 0$ which satisfy the frame theorems given in chapter 3 are selected to be $b \in (0.25, 1.875)$. Since we require that the redundancy in frames be kept to a minimum, it is important to keep the ratio of frame bounds B/A as close to one as possible. This factor, along with the number of dilation-translation nodes that can be allocated will influence the choice of b.

Chapter 4

Connection to Existing Methods

4.1 Regularization Theory

In this section we discuss regularization theory, and show how *symmetric* wavelets can be derived from the regularization framework of Poggio and Girosi [27].

Given a set of data $S = \{(x_i, y_i) \in \mathbf{R}^n \times R \mid i = 1, \cdots, J\}$, regularization theory gives the function f that minimizes a functional of the form

$$H(f) = \sum_{i=1}^{J} (y_i - f(x_i))^2 + \lambda \|Pf\|^2$$

Where λ is a positive real parameter called the regularization parameter, and P is an operator that captures whatever prior information on the smoothness of the function f is available. Using the Euler-Lagrange equations associated to the above problem, one obtains

$$f(x) = \sum_{i=1}^{J} c_i G(x; x_i) + p(x)$$

where the term $p(x)$ is a linear combination of functions spanning the null space of P, and arises because terms in the null space of P are invisible under the minimization of $H(f)$, and $G(x)$ is a Green's function of the operator $P^\dagger P$, with

$P^\dagger$ as the adjoint of P, and c_i is given by

$$c_i = \frac{(y_i - f(x_i))}{\lambda}.$$

Without looking for complicated operators one can think of the operation as a linear filtering operation which suppresses components in unwanted frequency bands. This fact is recognized in related literature (see, e.g.,[14]). For example, in order to arrive at Gaussian Radial Basis Functions(GRBF), one can consider the operation as passing f through a High-Pass Filter given by $e^{\frac{\|\omega\|^2}{\sigma}}$. As we will see later, the filtering function corresponds to $\frac{1}{\tilde{G}}$.

Since we are interested in obtaining wavelets as function approximators, it is relevant to consider whether wavelets can be derived within the regularization frame work. There are two motivations for this. One is that wavelet theory provides functions that range from *orthonormal bases* to *frames* and provides flexibility of choice with possible algorithmic improvements. The other is that although wavelet theory is now well-developed from the point of view of functional analysis and approximation theory, in many applications one confronts the problem of fitting a wavelet-based approximator on-line or off-line to an unknown system described only by the set of input-output data obtained from observations. In such problems, the relation between the number of data available and the number of parameters to be chosen in the wavelet approximator, the choice of the parameters either on-line or off-line etc., are not trivial issues. This is particularly appealing when the regular lattice structure used in wavelet theory causes an excessive number of wavelet units. Thus embedding wavelet theory in a regularization-statistical frame work is desirable.

First we will consider whether a mother wavelet can be derived from this

theory. The strong admissibility condition of a mother wavelet $\phi \in \mathbf{R}$ is

$$\int_{\mathbf{R}} \phi(x)dx = 0.$$

We have also shown in chapter 2 how to construct mother wavelets in $\mathbf{R}^n$ from a mother wavelet in $\mathbf{R}$. This suggests that we have to impose additional assumptions on P in the form of minimizing not only the high frequency energy, but also the energy in a small region ($B_\epsilon = \{\omega | -\epsilon < \|\omega\| < +\epsilon\}$) near zero-frequency.

One can include a weight for the constant term and a polynomial $p(x)$ without causing problems in practice. In fact, $p(x) = \sum_{j=1}^{n} w_j x_j$ is sometimes used to capture any linear dependencies that may exist.

For derivation, however, we look for a Band-Pass filtering function $\hat{B}(\omega)$ that vanishes at $\omega = 0$ and approaches zero as $\omega \to \infty$, such that $\frac{1}{\hat{B}(\omega)}$ provides the necessary filtering operation.

Then we can write the functional as

$$H(\hat{f}) = \sum_{i=1}^{J} \left(y_i - \int_{R^n} d\omega \hat{f}(\omega) e^{ix_i \cdot \omega} \right)^2 + \lambda \int_{R^n \setminus B_\epsilon} d\omega \frac{|\hat{f}(\omega)|^2}{\hat{B}(\omega)}$$

Minimizing the above functional with respect to $\hat{f}$ by setting the functional derivative to 0 under the limiting operation $\epsilon \to 0$ results in

$$\hat{f}(\omega) = \hat{B}(-\omega) \sum_{i=1}^{J} \frac{(y_i - f(x_i))}{\lambda} e^{ix_i \cdot \omega}.$$

The above result shows that for using the above theory in a general way, to approximate a function f, we have to assume a symmetric $B(\cdot)$ to have $\hat{B}(\omega)$ is real. Under this assumption, the Inverse Fourier Transform gives

$$f(x) = \sum_{i=1}^{J} c_i B(x - x_i) + p(x).$$

Here $p(x)$ and c_i are taken as defined earlier. As in the case of Gaussian Radial Basis functions, $p(x)$ is unnecessary since the null-space of the filtering operation is empty.

The restriction on symmetry rules out many orthonormal wavelet bases that are known in the literature. The Meyer wavelet and the Battle-Lemarié wavelet family appear to be the only known symmetric orthonormal wavelet bases; but it is also known that these wavelets have a large support; several compactly supported orthonormal wavelets bases are known, but all are non-symmetric [8].

Any symmetric wavelets (either frames or orthonormal bases) can be used. For example, taking $\hat{B}(\omega) = \|\omega\|^2 e^{-\frac{\|\omega\|^2}{2}}$ results in the Radial Mexican Hat:

$$\left(n - \|x\|^2\right) e^{\frac{-\|x\|^2}{2}}.$$

Tensor product of any symmetric wavelet can also be derived simply by considering $\hat{B}(\omega) = \hat{B}_1(\omega_1) \cdots \hat{B}_n(\omega_n)$, where $\hat{B}(\omega_i)$ are the 1-D Band-Pass Filters for $i = 1, \cdots, n$.

So far, we have considered the derivation of mother wavelets only. This results in an approximation scheme that uses *translations* only with a continuous wavelet transform. The essence of wavelet theory lies in the fact that it provides an elegant tool for spatio-spectral localization using dilations and translations. We will see that for the case considered above (i.e., symmetric mother wavelets), a *translation-dilation* structure with a continuous wavelet transform can be derived within the regularization framework. For this purpose, we combine two separate extensions for Hyper Basis Functions(HBF) [26].

- Assuming that the function is to have several levels of resolution, i.e., $f(x) = \sum_{m=1}^{M} f_m(x)$. Then, one can consider each level f_m at a dilation

level a^m in the forementioned procedure to arrive at

$$f(x) = \sum_{m=1}^{M} \sum_{i=1}^{J} c_{mi} B(a^m(x - x_i)).$$

- Weighted Norm. One considers the weighted function $B(\|x - x_i\|_W)$.

  Choose $W = \text{diag}\left(\mathrm{a^{m1}}, \cdots, \mathrm{a^{mN}}\right)$. This norming idea arises primarily as a means of taking into account the increased degrees of freedom that can result in dimensionality reduction.

Thus the basic approximation scheme can be advanced without the rigorous conditions of chapter 2, making it a suitable alternative when not using a regular lattice. It is important to note that while the original form of regularization theory fits a 'basis' corresponding to each data point, practical considerations require that approximate techniques be used to select a fewer number of basis elements [27]. Notice however that Regularization does not and cannot address the issue of sequential learning.

4.2 Radial Basis Functions

Radial Basis Functions provide another tool for function approximation. Although there exist several types of radial basis functions that can be used in approximations, the Gaussian Radial Basis Function (GRBF) of the form $G(x; x_i) = \frac{1}{\beta} e^{-\frac{\|x - x_i\|^2}{2\beta^2}}$ has been the subject of much attention because of its many desirable properties such as locality, separability, etc. A fixed value of β and a fixed sampling lattice for $x_i = k\Delta$ can be used to study the approximation properties given by

$$f = \sum_{k} c_k G(x; k).$$

One can use a multi-resolution scheme in which different levels of β are used. These functions however use a single parameter β to control all dimensions of x and are analogous to radial wavelets. In contrast, our focus is on tensor product wavelets. Therefore the idea of using weighted norm is relevant here. The weighting matrix is simply a diagonal matrix with the elements corresponding to different dimensions, i.e., the weighting matrix is $\text{diag}(\beta_1, \cdots, \beta_n)$. When such schemes are used, the result is essentially similar to our wavelet methodology. The only difference is that such structure is artificially imposed unlike the natural structure provided by wavelet theory, and hence strategies used in the choice of parameters may not have similar mathematical validity.

4.3 Projection Pursuit Regression(PPR)

Projection Pursuit [12] is an statistical technique that interprets high-dimensional data through well-chosen low-dimensional projections. In PPR, this technique is used in a successive refinement approach for non-parametric regression. Consider the single output case.

$$\hat{f} = \sum_{m=1}^{M} f_m \left({\alpha^T}_m x \right)$$

with

$$\alpha_m^T x = \sum_{j=1}^{p} \alpha_m^{(j)} x_j.$$

Here the f_m are single-valued ridge functions of a single variable. The parameters α_m^T as well as the functions f_m are chosen to simultaneously minimize the expected error. A forward stepwise procedure is used to select the model order M. It is clear such a strategy has many similarities to the wavelet methodology

when one considers the analogy between the m above and the dilation m. Indeed it is this observation that led to the matching pursuit (MP) algorithm [18] and later the orthogonal matching pursuit (OMP) [24], which orthogonalizes the 'basis' functions at each stage, just as the Orthogonal Least Squares (OLS) [3] orthogonalizes the Least Squares (LS) procedure.

When one considers off-line fitting of a model based on the complete set of observations and assumed levels of frequency content (and hence the dilations), one can construct a set of translation-dilation indices (the dictionary) from the data so that these data points lie in the 'receptive field' of the elements of the dictionary. Then it becomes possible to apply the OMP, and it is indeed the optimal strategy, but it will require more computations than the MP. But a central concern to us is that such a construction of dictionary requires so much *a priori* information, and is not practicable in many problems. Furthermore we place emphasis on on-line learning, and therefore faster methods. Our heuristic strategy is an essential tool that can be used directly to fit the model under this situation. In off-line, if optimality is a major concern, the dictionary automatically selected in this procedure can be further optimized by using OMP (normally simple strategies such as removing the dictionary elements with insignificant weights, also called 'wavelet shrinkage' can be used).

Chapter 5

Adaptive Control of Nonlinear Systems

We consider direct adaptive control of a SISO plant.

5.1 Problem Formulation

Here we limit our analysis to the class of non-linear systems that has a well established analytical framework in nonlinear control theory, namely, systems that have a canonical structure of the form:

$$x_1^{(n)}(t) = f(x_1(t), \dot{x}_1(t), \cdots, x_1^{(n-1)}(t)) + gu(t).$$

In general, $g = g(x_1, \dot{x}_1, \cdots, x_1^{(n-1)})$.

Assumptions:

We assume that $f(x)$ and $g(x)$ are sufficiently smooth, that $g^{-1}(x)$ exists (or $|g(x)| \geq b_g > 0$) and is smooth in the region of our interest; the assumption on $g^{-1}(x)$ implies that $g(x)$ is of the same sign everywhere in the region, and without loss of generality we can take it as positive. It is also assumed that $|f(x)|$

and $|g(x)|$ are upper bounded in the region of interest by known functions $M_f(x)$ and $M_g(x)$. Measurability of the state vector x, and Persistency of Excitation are assumed (this latter point will be discussed later). It is emphasized that unless otherwise stated, no other information about the functions is assumed.

More general classes of systems can only be solved under certain more restrictive assumptions or on an *ad hoc* basis.

It is well known that many practical control problems such as robotic manipulator control can be reduced to the above canonical form (see, e.g., [31]); neural adaptive control schemes for plant models of this form or minor variants thereof have been studied by Chen and Khalil [3] using Backpropagation (Hyperbolic tangent activation functions) neural networks, and by others [30, 28] using Gaussian Radial Basis Functions.

Let $x = (x_1, \dot{x}_1, \cdots, x_1{}^{(n-1)})^T$, $x_d = \left(x_{1d}, \dot{x}_{1d}, \cdots, x_{1d}^{(n-1)}\right)^T$ be the state and desired vectors respectively. If $f(x)$ and $g(x)$ are completely known, we can consider

$$u(t) = g^{-1}(x_{1d}^n(t) + u_{pd}(t) - f(x)), \tag{5.1}$$

where $u_{pd}(t) = -k^T e(t)$, with $e(t) = x - x_d$.

The problem is that $f(x)$ and $g(x)$ are unknown (except for the assumptions made earlier).

Hence we have to make suitable approximations to the unknown functions through interactions with the plant. It is in this context that neural networks find their usefulness in control. Wavelet-based approximation networks are yet another way of providing the required approximation capability. Let $\hat{f}$ and $\hat{g}$ be the approximation provided by such a network. The $u(t)$ in 5.1 is redefined as,

$$u(t) = \hat{g}(x)^{-1}(x_{1d}^n(t) + u_{pd}(t) - \hat{f}(x)). \tag{5.2}$$

Upon substituting the control law in 5.2, the error equation becomes,

$$\dot{e} = Ae + b(-\hat{f} + f(x)) + b(g(x) - \hat{g}(x))u,$$

where

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 1 \\ -k_1 & -k_2 & -k_3 & \cdots & -k_n \end{pmatrix}$$

and

$$b = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

This shows that if $\hat{f}(x)$ and $\hat{g}(x)$ continuously track the unknown functions $f(x)$, and $g(x)$ respectively, while maintaining boundedness of the error e (and hence boundedness of the state vector), the control problem is solved by taking the gain of the PD controller $k = (k_1, \cdots, k_n)$ to make a Hurwitz matrix.

In practice we can only attempt to guarantee an approximation in the form

$$|\hat{f}(x) - f(x)| \leq \epsilon_{af}$$

and

$$|\hat{g}(x) - g(x)| \leq \epsilon_{ag},$$

where $\epsilon_{af}, \epsilon_{ag}$ are the uniform upper bounds on the error in approximating

$f(x)$ and $g(x)$ by

$$\hat{f}(x) = \hat{\alpha}_f + \sum_{m,n} \hat{c}_{mn}\psi_{mn}$$

and

$$\hat{g}(x) = \hat{\alpha}_g + \sum_{m,n} \hat{d}_{mn}\psi_{mn}$$

respectively. Here ψ_{mn} are the wavelet frames, and we include α_f and α_g so that they can capture the mean values of f and g, if non-zero. Such terms are not necessary in approximations based on RBFN or sigmoidal feed-forward networks.

Moreover, in practice an error results from mis-tuning of the parameters $\hat{\alpha}_f, \hat{\alpha}_g, \hat{c}_{mn}, \hat{d}_{mn}$. Let the actual values of the parameters be $\alpha_f, \alpha_g, c_{mn}, d_{mn}$ respectively, and let the functions they constitute be $\tilde{\hat{f}}, \tilde{\hat{g}}$. Then in order to consider this error we define the following:

$$e_{c_{mn}} = -c_{mn} + \hat{c}_{m,n}$$

$$e_{d_{mn}} = -d_{mn} + \hat{d}_{m,n}$$

$$e_{\alpha_f} = -\alpha_f + \hat{\alpha}_f$$

$$e_{\alpha_g} = -\alpha_g + \hat{\alpha}_g$$

We have to consider the resulting effect on the linearized equation and show that the presence of these mis-match error terms does not lead to instability during adaptation and learning.

5.2 Stability

We have

$$\dot{e}(t) = Ae(t) + b\left(e_{\alpha_f} + \sum_{m,n} e_{c_{mn}}(t)\psi_{mn}(t)\right)$$

$$+bu(t)\left(e_{\alpha_g}+\sum_{m,n}e_{d_{mn}}(t)\psi_{mn}(t)\right)+dist(t)b, \tag{5.3}$$

where $dist(t)=f(x)-\hat{\alpha}_f-\sum_{m,n}\hat{c}_{mn}\psi_{mn}+\left(g(x)-\hat{\alpha}_g-\sum_{m,n}\hat{d}_{mn}\psi_{mn}\right)u$.

First we consider the case where g(x) is a **known constant**; without loss of generality we can take $g(x)=1$.

Case: g(x)=1 Then 5.2 becomes

$$u=x_{1d}^{n}(t)+u_{pd}(t)-\tilde{\hat{f}}(x). \tag{5.4}$$

For the moment, assuming that the network has sufficient units (this point will be elaborated later in this chapter) to approximate f with the uniform error bound ϵ_{af} for all practical values of the state vector, we have that

$$|dist(t)|\leq\epsilon_{af}.$$

There exist different adaptive control approaches and corresponding approaches to ensure stability. We follow Lypunov design methods, whereby the adaptation law derived is consistent with Lyapunov stability.

Since A is strictly Hurwitz, by the Kalman-Yakubovich-Popov lemma, there exist symmetric and positive definite matrices P and Q such that

$$PA+A^TP=-Q$$

Therefore we can consider the Lyapunov function

$$V(e,e_{c_{mn}},e_{\alpha_f})=\frac{1}{2}e^TPe+\frac{1}{2k_f}\left(\sum_{m,n}e_{c_{mn}}^2+e_{\alpha_f}^2\right)$$

where k_f is a suitable positive gain value in adaptation.

$$\dot{V}(e,e_{c_{mn}},e_{\alpha_f})=-\frac{1}{2}e^TQe+e^TPb\left(f-\tilde{\hat{f}}\right)+\frac{1}{k_f}\left(\sum_{m,n}e_{c_{mn}}\dot{e}_{c_{mn}}+e_{\alpha_f}\dot{e}_{\alpha_f}\right) \tag{5.5}$$

The first term is non-positive. Define $s \triangleq e^T Pb$, the augmented error. The second term can be made zero by considering a suitable adaptation law for c_{mn} and α_f. Looking at 5.5, we want to cancel out the second and third terms. Let us consider the adaptive laws,

$$\dot{e}_{c_{mn}} = -k_f s \psi_{mn}$$

and

$$\dot{e_{\alpha f}} = -k_f s.$$

By definition of the parameter error terms, we have equivalently,

$$\dot{c}_{mn} = k_f s \psi_{mn},$$

and

$$\dot{\alpha_f} = k_f s.$$

From 5.5 we arrive at

$$\dot{V}(e, e_{c_{mn}}, e_{\alpha_f}) = -\frac{1}{2} e^T Q e + s e_f,$$

where e_f is the instantaneous error inherent in approximating f by $\hat{f}$, i.e., $f - \hat{f}$, and is upper bounded by ϵ_{af}. The presence of this term necessitates some modifications to the control law and adaptive laws. Let us consider a modification to the control law by adding a new term to u in 5.4 in the form $u_\Delta = -sgn(s)\Delta_f$ where Δ_f is a positive value chosen such that $\Delta_f > \epsilon_{af}$. From 5.4, the new control input is given by

$$u = x_{1d}^n(t) + u_{pd}(t) - \tilde{\hat{f}}(x) + u_\Delta.$$

Then

$$\dot{V}(e) = -\frac{1}{2}e^T Qe + s(e_f - sgn(s)\Delta_f).$$

We see that the second term is forced to be non-positive by means of the fact $|e_f| < \epsilon_{af} < \Delta_f$. Hence, initial boundedness of the state vector and the parameters implies that they remain bounded for all time. By a simple application of Barbalat's lemma (see A.4), asymptotic convergence of the tracking error vector is established.

Since we allow for the possibility of large errors in approximation, it is desirable to have a reasonably large Δ_f. We use a dead-zone d, i.e., we adapt c_{mn} and α_f only if $e^T Pe > d^2$ and $\dot{c}_{mn}, \dot{\alpha}_f$ are 0 otherwise.

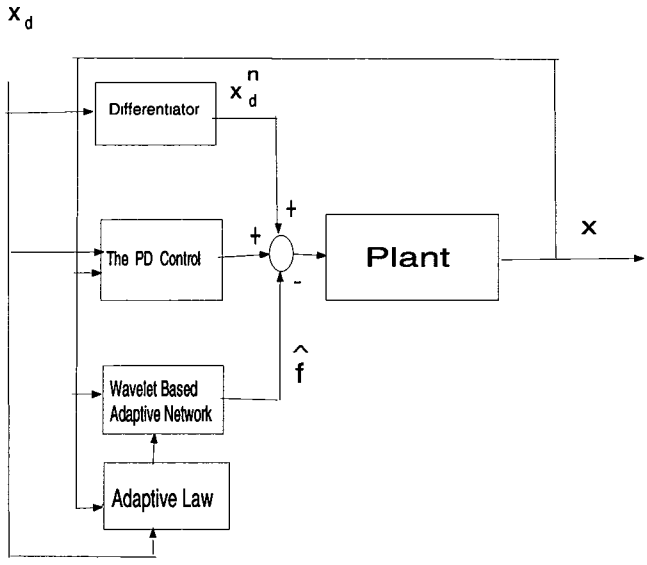


Figure 5.1: The Control Architecture

Case: g(x) is unknown

The ψ_{mn} are normalized to unit maximum amplitude rather than to unit norm.

Therefore,

$$|\hat{f}| \leq |\alpha_f| + \sum_{m,n} |c_{mn}|,$$

and

$$|\hat{g}| \leq |\alpha_g| + \sum_{m,n} |d_{mn}|.$$

Thus $\hat{f}$ and $\hat{g}$ are upper bounded; also $\hat{g}$ is ensured to be invertible by enforcing a lower bound to the network output during adaptation.

Hence the control law is,

$$u(t) = \tilde{\hat{g}}(x)^{-1}(x_{1d}^n(t) + u_{pd}(t) - \tilde{\hat{f}}(x)) \tag{5.6}$$

To ensure that the adaptation laws are consistent with Lyapunov stability, consider the Lypunov function

$$V(e, e_{c_{mn}}, e_{d_{mn}}) = \frac{1}{2}e^T P e + \frac{1}{2k_f}\left(\sum_{m,n} e_{c_{mn}}^2 + e_{\alpha_f}^2\right) + \frac{1}{2k_g}\left(\sum_{m,n} e_{d_{mn}}^2 + e_{\alpha_g}^2\right),$$

where k_f and k_g are suitable positive adaptation gains.

$$\begin{aligned} \dot{V}(e, e_{c_{mn}}, e_{d_{mn}}) &= -\frac{1}{2}e^T Q e + e^T P b\left(f - \tilde{\hat{f}} + u\left(g - \tilde{\hat{g}}\right)\right) \\ &\quad + \frac{1}{k_f}\left(\sum_{m,n} e_{c_{mn}}\dot{e}_{c_{mn}} + e_{\alpha_f}\dot{e}_{\alpha_f}\right) \\ &\quad + \frac{1}{k_g}\left(\sum_{m,n} e_{d_{mn}}\dot{e}_{d_{mn}} + e_{\alpha_g}\dot{e}_{\alpha_g}\right) \end{aligned} \tag{5.7}$$

This suggests that we can attempt the following adaptation laws (with s as as the augmented error defined earlier):

$$\dot{c}_{mn} = k_f s \psi_{mn},$$

$$\dot{\alpha}_f = k_f s,$$

$$\dot{d}_{mn} = k_g u s \psi_{mn},$$

and

$$\dot{\alpha}_g = k_g u s.$$

The presence of u in the last two laws should be noted.

Upon substituting these laws in 5.7 we get

$$\dot{V}(e) = -\frac{1}{2}e^T Q e + s e_f + s u e_g,$$

where e_f and e_g are the disturbances caused by the inherent error in approximating f by $\hat{f}$, i.e., $e_f = f - \hat{f}$ and g by $\hat{g}$, i.e., $g - \hat{g}$ respectively. The presence of these terms again lead to modifications similar to the case of $g(x) = 1$.

Consider the additional control term

$$u_\Delta = -sgn(s)\frac{\Delta_f}{b_g(x)} - sgn(s)u\frac{\Delta_g}{b_g(x)}.$$

From 5.6, the new input u is given by

$$u = u_0 + u_\Delta,$$

where

$$u = \tilde{\hat{g}}(x)^{-1}(x_{1d}^n(t) + u_{pd}(t) - \tilde{\hat{f}}(x)).$$

Then the u appearing in the adaptive laws d_{mn} and α_g will be changed to u_0, i.e.,

$$\dot{d}_{mn} = k_g u_0 s \psi_{mn},$$

and

$$\dot{\alpha}_g = k_g u_0 s.$$

Then

$$\dot{V}(e) = -\frac{1}{2}e^T Q e + s\left(e_f - sgn(s)g(x)\frac{\Delta_f}{b_g}\right) + su\left(e_g - sgn(s)g(x)\frac{\Delta_g}{b_g}\right).$$

By making $|e_f| < \epsilon_{af} < \Delta_f$ and $|e_g| < \epsilon_{ag} < \Delta_g$ we see that $\dot{V}(e)$ is ensured to be non-positive. This ensures boundedness of the state vector and the parameters. Again, Barbalat's lemma can be used to show (see appendix A.4) that asymptotic tracking is established.

5.3 Effects Due to Dynamic Model Selection

If we are to attempt on-line learning of the structure in addition to on-line adaptation of the weights, it is imperative that we take into account the effects of incomplete set of indices (m, n) leading to violation of the upper bounds on approximation given by ϵ_{af} and ϵ_{ag}. This problem can be solved by taking Δ_f and Δ_g initially large and then gradually reducing them as a function of the error vector e. This has the merit of ensuring that energy is not wasted unnecessarily. Such as an adjustment is made possible by the fact that although the error vector in itself does not provide any information on the dilation-translation indices to be learnt, this vector can be used in combination with the state vector to give useful information on the proximity to existing translations, and dilations.

A crucial limitation of any fixed or adaptive control schemes without a 'learning' component is that unmodelled dynamics cannot be accounted for. For instance, in the case of robotic manipulator control, the friction terms are difficult to model accurately, whereas the inertia terms are well modelled. Thus both adaptive and fixed control strategies that are based on an assumed plant model (with either known or unknown functions) are difficult to control. This has led some researchers to propose learning control schemes when the action to be performed is repetitive. In [6], Craig proposed a linear filter based learning in combination with a fixed controller on the assumption that the fixed controller

can provide sufficient control so that the resulting error without the learning component is small. Such schemes have limited capability and cannot be used in more general plants. Neural networks can be used in place of the linear filter. Such an scheme can in principle combine adaptation and learning. In using non-local networks, all weights get adjusted at each novel situation (in the worst case at each iteration) and learning is essentially forgotten. What is required is a scheme that captures the underlying function *and* adapts to novel situations as well; i.e., a scheme that effectively adapts to novel situations without erasing the learnt portion. To some extent, local learning has the capability of achieving this since at every iteration only those weights corresponding to a local region of the input space get adjusted. However, after a small period of operation, effective learning may be lost since adaptation could have occurred virtually in the whole range of operation and erased any learning. This leads us to consider schemes that combine adaptation and learning in the sense described above. Our proposed scheme is indicated for further work in figure 5.2. Some aspetcs of this can be seen in the work of Farrel, et al. [10].

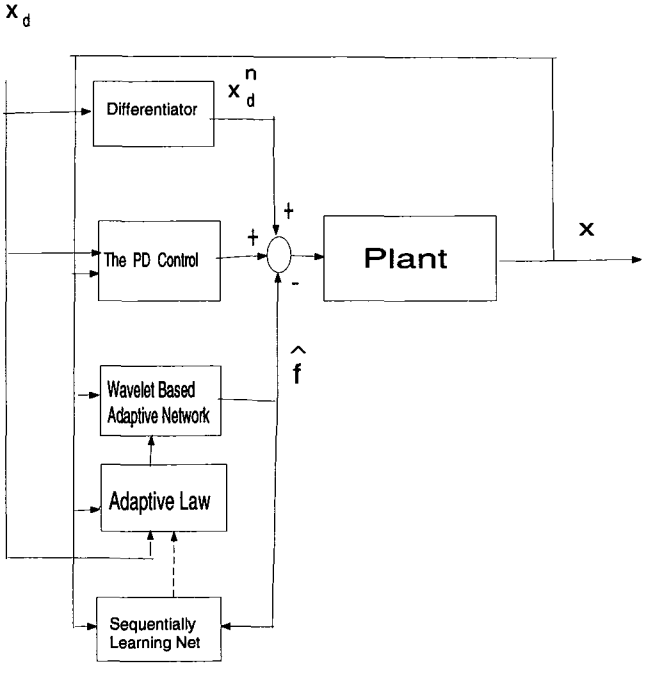


Figure 5.2: Proposed Control Architecture

Chapter 6

Simulation Results and Conclusions

In this chapter we present some simulation results. These simulations should not be interpreted as final; rather we present these as an indication of the merits of the procedures we had explained in earlier chapters, and will be improved in future work.

6.1 Simulation for learning algorithms

6.1.1 One-dimensional problems

The following function [34] was used.

$$f(x) = \begin{cases} -2.186x - 12.864 & -10 \leq x < -2 \\ 4.246x & -2 \leq x < 0 \\ 10e^{-0.05x-0.5}\sin\left[(0.03x + 0.7)\right] & 0 \leq x \leq 10 \end{cases}$$

These results in figures 6.2, 6.3 clearly show the success of our methodology in the one dimensional case. The performance with 22 units in this case shows how our learning algorithm automatically adds nodes where more nodes are necessary (high-frequency variations), and its significance lies in the fact that we use an

on-line sequential method (i.e., data are presented only once sequentially, and no non-linear optimization procedures used). As is expected in any sequential method, we needed roughly eight to ten times more data than off-line methods. This should not be a problem since each incoming data point can be discarded after presentation to the network. Figure 6.4 shows the network obtained for a reduced absolute error. The required generalization performance was very adequate in this case. This resulted in 39 nodes and a MSE of 0.005.

In another simulation, the Hermite Function $f(x) = 1.1(1-x+2x^2)\exp\{-\frac{x^2}{2}\}$ is used as in [17].

6.2 Simulations for Adaptive Control

For simulating the plant models considered in chapter 5, we consider the following two-dimensional plant function $f(x)$ as in [30]

$$f(x) = -4\left(\frac{\sin(4\pi x_1)}{\pi x_1}\right)\left(\frac{\sin(\pi \dot{x_1})}{\pi \dot{x_1}}\right)^2,$$

and $g(x) = 1$.

The plant output is taken as $y = x_1$.

The fixed network structure was attempted. A total of 35×35 nodes were required for f and g. The simulations were run using C.

The following values are selected: $k_1 = 1, k_2 = 20$. Then we obtain

$$A = \begin{pmatrix} 0 & 1 \\ -1 & -20 \end{pmatrix}.$$

Thus

$$P = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

satisfies the conditions inherent in Kalman-Yakubovich-Popov lemma. Hence

$$Pb = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

and $s = e^T Pb = e_1 + 2\dot{e}_1$.

Figures 6.8, 6.9, 6.10 show the tracking performance.

6.3 Conclusions

The use of multi-dimensional wavelet theory in network construction was established, and the method of using tensor-product wavelets was theoretically and experimentally studied. A methodology for building the network sequentially on-line was developed and shown to give good results in one-dimension. In high dimensional problems, for the same degree of success, we have to work with an inordinate amount of data, but we emphasize that this should not cause concern since the methodology is on-line.

Our results demonstrate the feasibility of using wavelets as network functions in adaptive control situations. While having certain similarities with [30]in the overall methodology used, our scheme differs from their methods in a number of ways. We do not use the sliding scheme and we implement the network for g rather than g^{-1}. Using wavelet network over the Gaussian network does reduce the number of nodes, but not sufficiently enough to make local learning practicable for high dimensional problems. Feedforward neural networks used by other researchers (see for e.g., [3]) result in local stability, with the condition that initial errors be small. Our method has no such restrictions (only boundedness is required), but this generally holds for all networks that have

a linear-in-the parameter structure. The non-local approximation property of feed-forward networks enables fewer network units to be used.

6.4 Future Directions

Developing sequentially growing more compact networks needs to be studied further from a statistical framework. Research in the direction of obtaining sample complexity estimates and generalization bounds for such compact networks is indicated. Recent work by Niyogi and Girosi [22] contains good results for RBFNs, etc, and the references contained therein are good sources on this problem. The simple idea of applying a threshold to the wavelet coefficients is given more theoretical analysis by Donoho and Johnstone [9] for orthogonal wavelets in the context of estimation theory and it should prove worthwhile to investigate the underlying connections between general (non-orthogonal) wavelet networks and non-parametric estimation further. Also, it will be useful to advance non-uniform sampling techniques to irregular lattice-based wavelet theory.

For control, further experimentation is needed on the practicality of implementing the adaptive/learning approach to complex systems. In a recent paper Narendra, et al. [21] have suggested broadening the simulations to multivariable systems, and report that attempting simultaneous identification and control leads to unstable results. Moreover, theoretical results on such problems as "persistency of excitation" in the adaptive/learning case, and implementing control schemes for more general plant models (for example, using recurrent networks) are also desirable. Considering that a plethora of existing methods (various neural schemes, CMAC, RBFN,WN, fuzzy systems, other hybrid learning

schemes) are scattered in the literature without any clear indication of comparative merits, comparative theoretical and experimental studies to unambiguously determine under what general conditions one method outperforms another are also necessary.

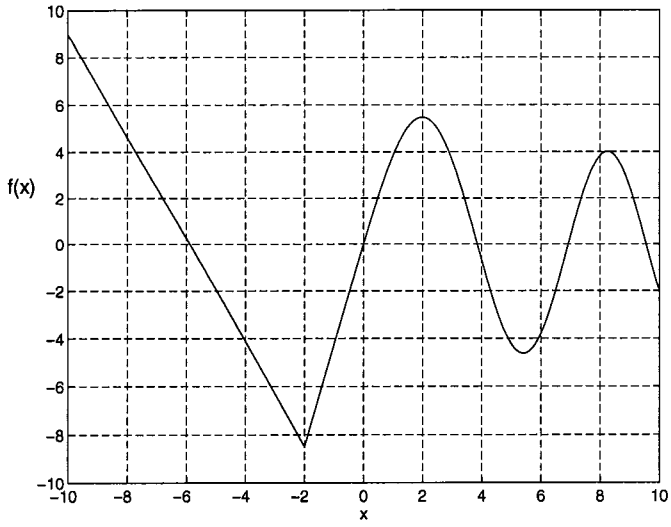


Figure 6.1: The Piece-wise Continuous Function

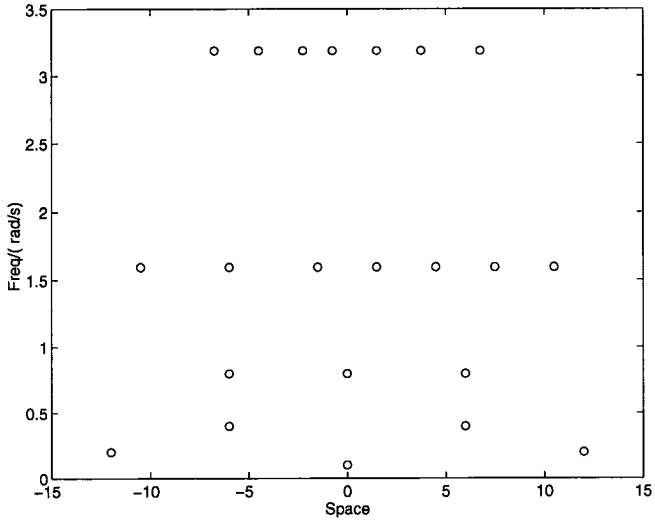


Figure 6.2: The Sequentially Learnt Structure of The Network When The Maximum Absolute Generalization Error is 0.6

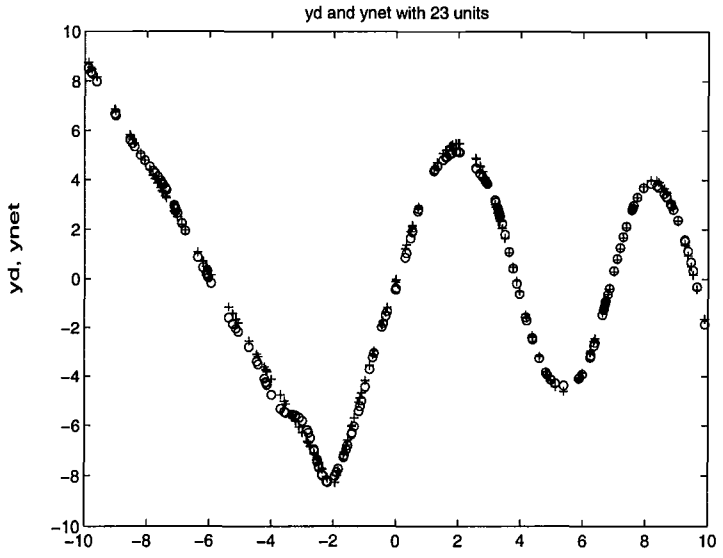


Figure 6.3: The Generalization Performance on a Test Set of 200 Data. yd('+') and ynet('o') are the Actual Function and Network Output Respectively

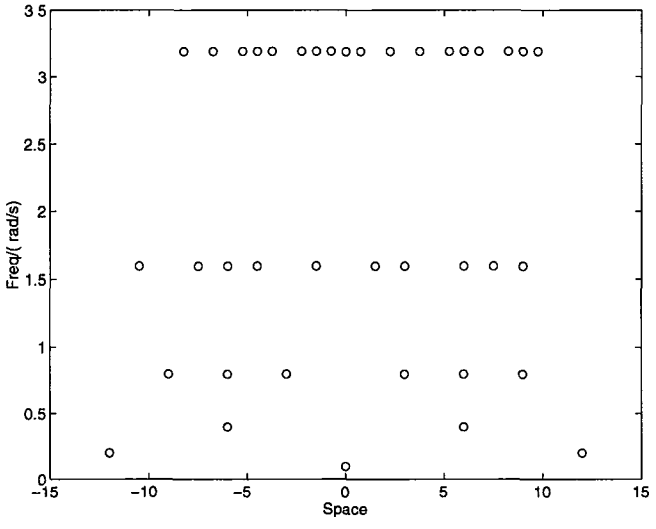


Figure 6.4: The Sequentially Learnt Structure of the Network When the Maximum Absolute Generalization Error is 0.2

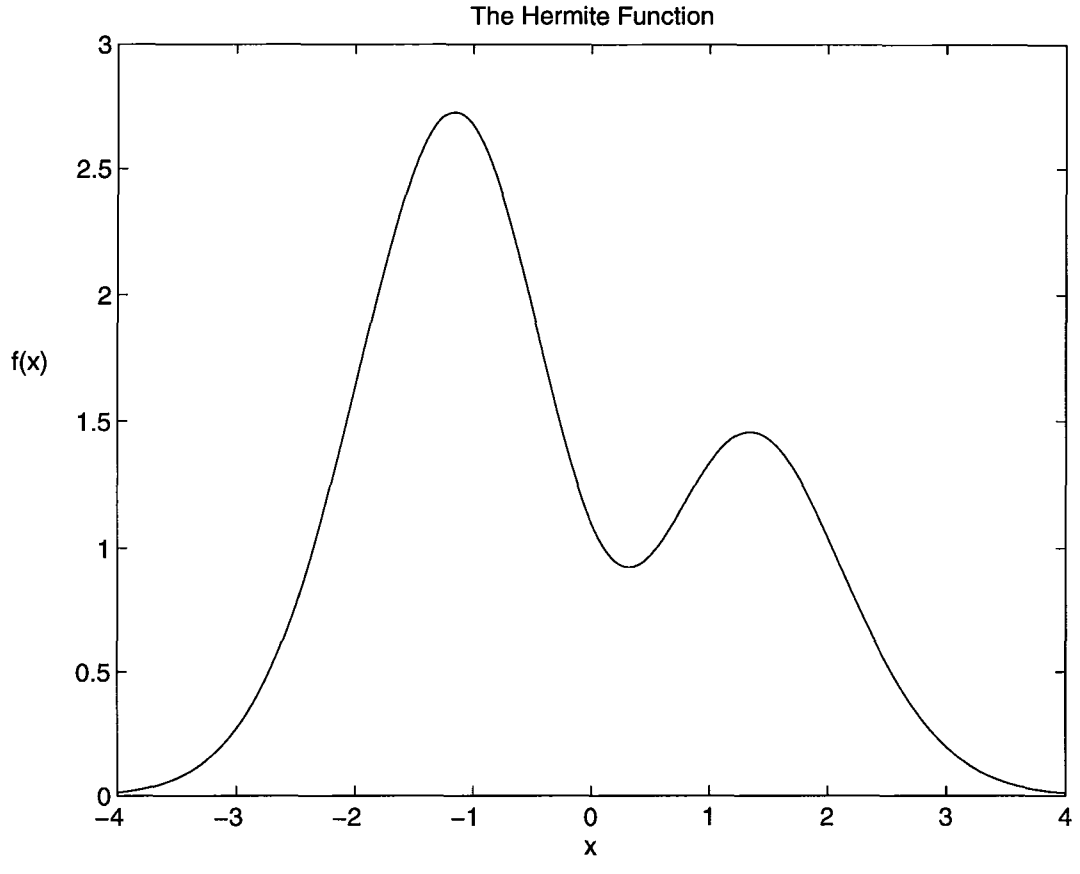


Figure 6.5: The Hermite Function

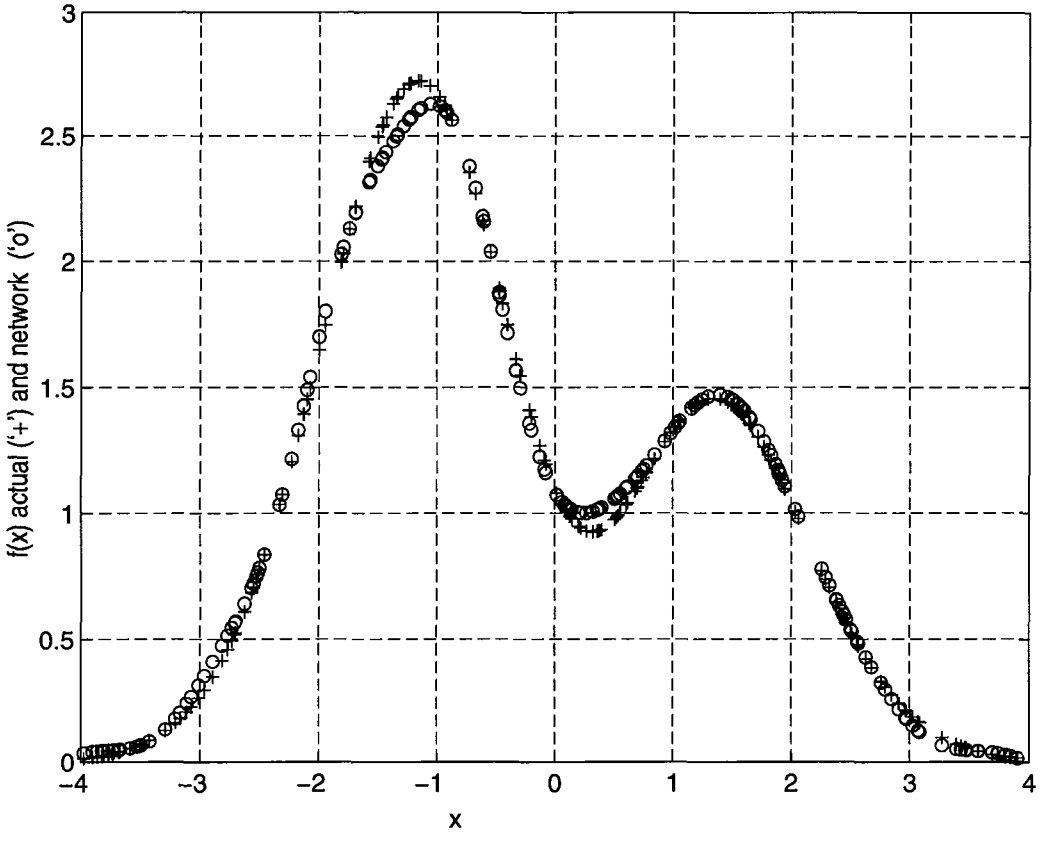


Figure 6.6: The Generalization Performance: '+' and 'o' Mark the Actual Function and Network Output Respectively

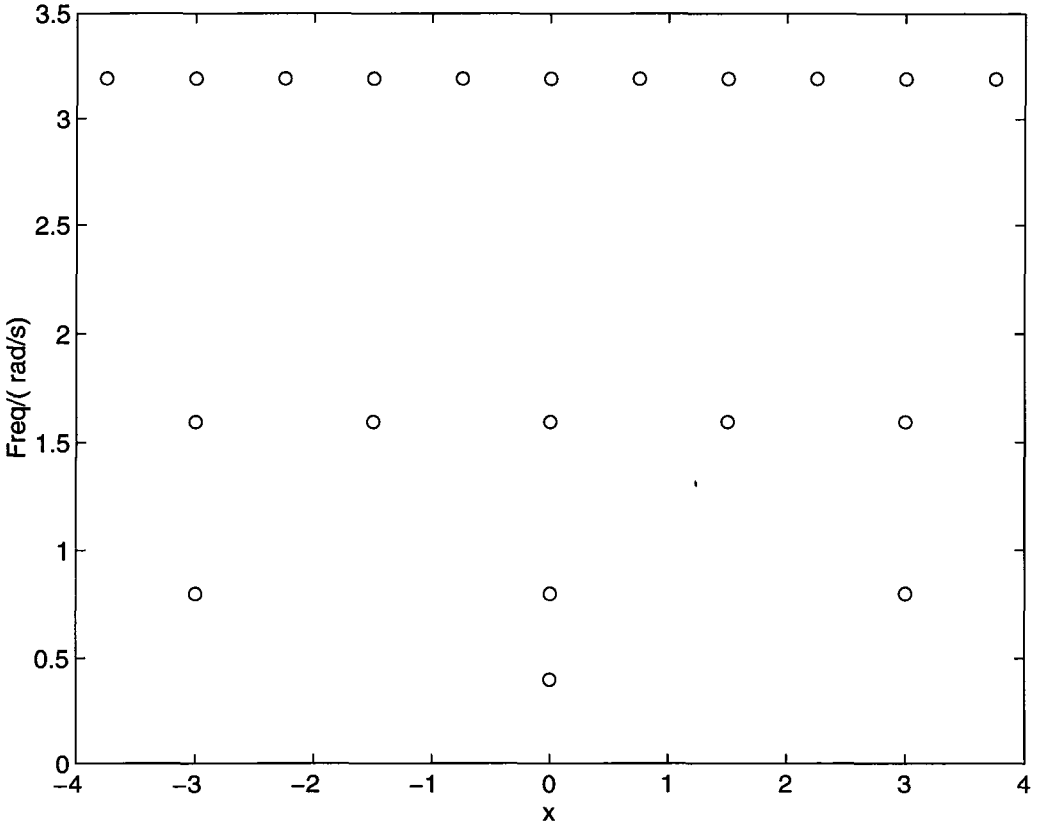


Figure 6.7: The Sequentially Learnt Lattice Structure

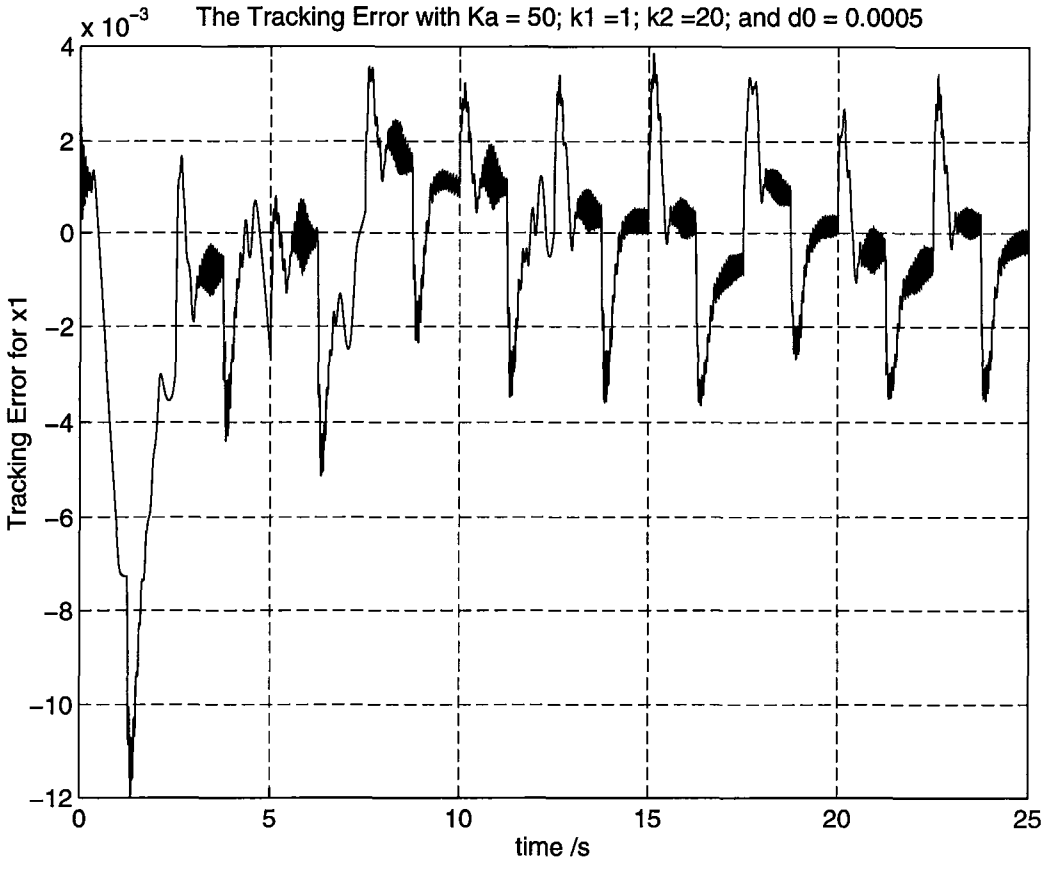


Figure 6.8: The Tracking Error

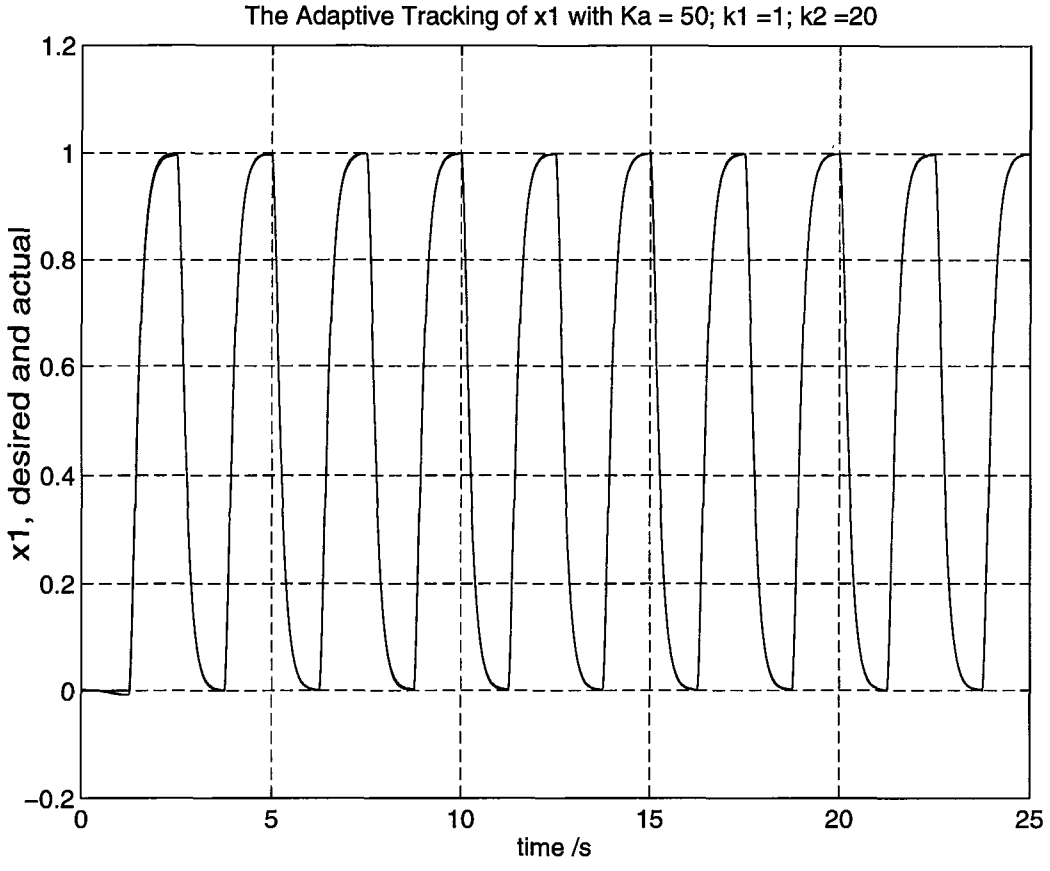


Figure 6.9: The Tracking Performance

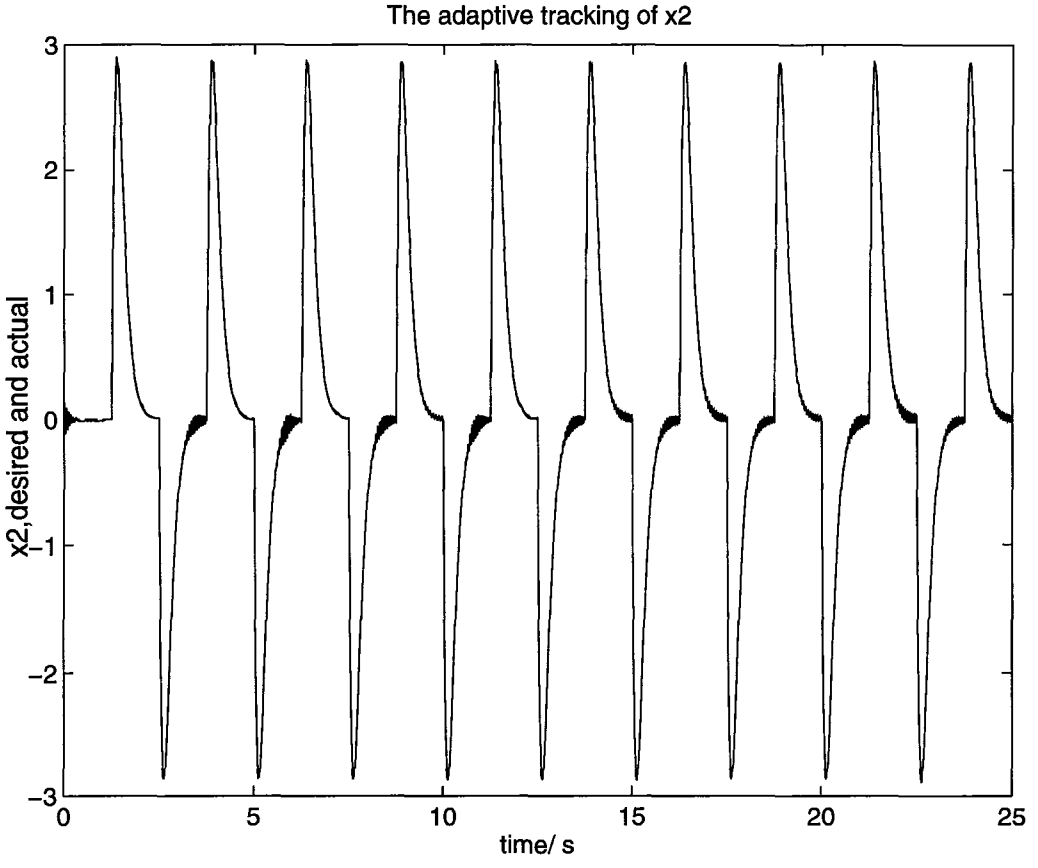


Figure 6.10: The Tracking Performance

Appendix A

A.1 Proof of theorem 1

As our proof closely follows Daubechies' scheme for 1-D (see [7] and the section 3.3.2. of [8]), we expand on these results to show the validity of the conditions given to many dimensions.

First, we need the following generalization of the Poisson formula:

$$\sum_{k\in Z^n} e^{iCk^T x} = \left(\frac{2\pi}{C}\right)^n \prod_{j=1}^{n} \sum_{k_j\in Z} \delta\left(x_j - \frac{2\pi}{C}k_j\right)$$

where i is the imaginary unity and C is any real non zero constant. It can be verified by simple computations.

By applying this generalized Poisson formula and the Parseval's theorem, straight forward computations give

$$\sum_{l,k} |\langle \psi_{l,k}, f\rangle|^2 \;=\; \left(\frac{2\pi}{b}\right)^n \sum_{l} \int \mathrm{d}\omega\, |\widehat{\psi}(a^l\omega)|^2 \cdot |\widehat{f}(\omega)|^2 + \Lambda,$$

where

$$\Lambda = \left(\frac{2\pi}{b}\right)^n \sum_{l} \sum_{k\neq 0} \int \mathrm{d}\omega\, \widehat{\psi}(a^l\omega)\, \overline{\widehat{\psi}\left(a^l\omega - \frac{2\pi}{b}k\right)}\, \overline{\widehat{f}(\omega)}\, \widehat{f}\left(\omega - \frac{2\pi}{a^l b}k\right),$$

and $k \neq 0$ means *at least one component* of k is not zero.

By applying the Cauchy-Schwarz inequality, we get

$$|\Lambda| \leq \left(\frac{2\pi}{b}\right)^n \sum_{k\neq 0} \left[\beta\left(\frac{2\pi}{b}k\right)\beta\left(-\frac{2\pi}{b}k\right)\right]^{\frac{1}{2}} \|f\|^2$$

where $\beta(\cdot)$ is as defined in (2.5). This inequality together with the three conditions of the theorem gives

$$\left(\frac{2\pi}{b}\right)^n \left\{m(\psi,a) - \sum_{k\neq 0}\left[\beta\left(\frac{2\pi}{b}k\right)\beta\left(-\frac{2\pi}{b}k\right)\right]^{\frac{1}{2}}\right\}\|f\|^2 \leq \sum_{l,k}|\langle\psi_{l,k},f\rangle|^2$$
$$\leq \left(\frac{2\pi}{b}\right)^n \left\{M(\psi,a) + \sum_{k\neq 0}\left[\beta\left(\frac{2\pi}{b}k\right)\beta\left(-\frac{2\pi}{b}k\right)\right]^{\frac{1}{2}}\right\}\|f\|^2.$$

The only thing left is to verify that condition (2.4) ensures the convergence of the multi-indexed series

$$\sum_{k\neq 0}\left[\beta\left(\frac{2\pi}{b}k\right)\beta\left(-\frac{2\pi}{b}k\right)\right]^{\frac{1}{2}},$$

and implies that the sum tends to zero when $b \to 0$, so that the coefficients of $\|f\|^2$ in the above inequalities are strictly positive for small enough b.

By (2.4) we have,

$$\beta(\eta) \leq C_\epsilon\left(1+\eta^T\eta\right)^{-\frac{n(1+\epsilon)}{2}}.$$

This leads to

$$\sum_{k\neq 0}\left[\beta\left(\frac{2\pi}{b}k\right)\beta\left(-\frac{2\pi}{b}k\right)\right]^{\frac{1}{2}} \leq C_\epsilon\left(\frac{b}{2\pi}\right)^{n(1+\epsilon)} \sum_{k\neq 0} \left[\left(\left(\frac{b}{2\pi}\right)^2 + |k_1|^2 + \cdots + |k_n|^2\right)^n\right]^{-\frac{1+\epsilon}{2}}.$$

Considering the inequality

$$\left(C + |k_1|^2 + \cdots + |k_n|^2\right)^n \geq \left(C+|k_1|^2\right)\left(C+|k_2|^2\right)\cdots\left(C+|k_n|^2\right)$$

where C is any positive constant, we see that the series in k converges. Moreover, as $b \to 0$, this sum tends to zero. The proof of the theorem is thus established.

□

A.2 Proof of Theorem 2

As in the proof of theorem 1, we use the Poisson formula in n dimensions, and the Parseval's Theorem in relation to Fourier Transforms. The steps involved are similar to the proof of theorem 1 as shown below.

We arrive at

$$\sum_{j,k} |\langle \psi_{j,k}, f \rangle|^2 = (2\pi \ \det T^{-1})^n \sum_j \int \mathrm{d}\omega |\widehat{\psi}(D_{-j}\omega)|^2 |\widehat{f}(\omega)|^2 + \Lambda$$

where

$$\begin{aligned} \Lambda &= (2\pi \ \det T^{-1})^n \sum_j \sum_{|k|\neq 0} \int \mathrm{d}\omega \widehat{\psi}(D_{-j}\omega) \overline{\widehat{\psi}(D_{-j}w - 2\pi T^{-1}k)} \\ &\quad \overline{\widehat{f}(\omega)} \widehat{f}(\omega - 2\pi D_{-j} T^{-1}k) \end{aligned}$$

The third condition of the theorem implies the decay of β as

$$\beta(\eta) \leq (1 + \eta^T \eta)^{-\frac{n(1+\epsilon)}{2}} \cdot C_\epsilon.$$

Hence

$$\begin{aligned} \sum_{|k|\neq 0} \left[\beta(2\pi T^{-1}k)\beta(-2\pi T^{-1}k)\right]^{\frac{1}{2}} &< C_\epsilon \left(\frac{\det T}{2\pi}\right)^{n(1+\epsilon)} \sum_{|k|\neq 0} \\ &\quad \left[\left(\left(\frac{\det T}{2\pi}\right)^2 + k^T k\right)^n\right]^{-\frac{1+\epsilon}{2}}. \end{aligned}$$

The multi-indexed series converges as in theorem 1. Moreover, it is easily seen that when $b_i, i = 1, \cdots, n \to 0$, the sum $\to 0$; and the limit on b is given by

$$b_c = \inf \left\{ b | m(\psi, a) \leq \sum_{|k|\neq 0} \left[\beta(2\pi T^{-1}k)\beta(-2\pi T^{-1}k)\right]^{\frac{1}{2}} \right\}.$$

Again, the bounds and inequalities are considered element-wise.

This completes the proof of the theorem. □

A.3 The relations required in section 3.4.1

Let j denote the index of one dimension, i.e., $j \in \{1, \cdots, N\}$. Consider the translations x_{ji} for $i \in \{1, \cdots, n-1\}$ and x_{jn} at the same dilation level $m_j = m$. We have

$$\psi_{ij} = \left(1 - a^{2m}\left(x_j - x_{ji}\right)^2\right) e^{-\frac{a^{2m}\left(x_j - x_{ji}\right)^2}{2}}.$$

Then

$$\begin{aligned} < \psi_{ij}, \psi_{nj} > \; = \; & \int_{-\infty}^{\infty} \left(1 - a^{2m}\left(x_j - x_{ji}\right)^2\right) e^{-\frac{a^{2m}\left(x_j - x_{ji}\right)^2}{2}} \\ & \left(1 - a^{2m}\left(x_j - x_{jn}\right)^2\right) e^{-\frac{a^{2m}\left(x_j - x_{jn}\right)^2}{2}} \mathrm{d}x_j. \end{aligned}$$

By writing out this expression, and using the fact that

$$\|\psi_{ij}\|^2 = \frac{3}{4}\sqrt{\pi}a^{-m}$$

we arrive at the following result.

$$< \psi_{ij}, \psi_{nj} >= e^{-\frac{a^{2m}d_j^2}{4}} \left(1.0 - a^{2m}d_j^2 + \frac{1}{12}a^{4m}d_j^4\right)$$

where

$$d_j = x_{ji} - x_{jn}.$$

Therefore, zero-crossings occur at

$$\left(a^m d - 6.0\right)^2 = 24.0,$$

i.e., $|d| = 1.0493 * a^{-m}$, or$3.3014 * a^{-m}$. Working within the framework of a regular lattice, we may choose to have

$$|d| = a^{-m}b.$$

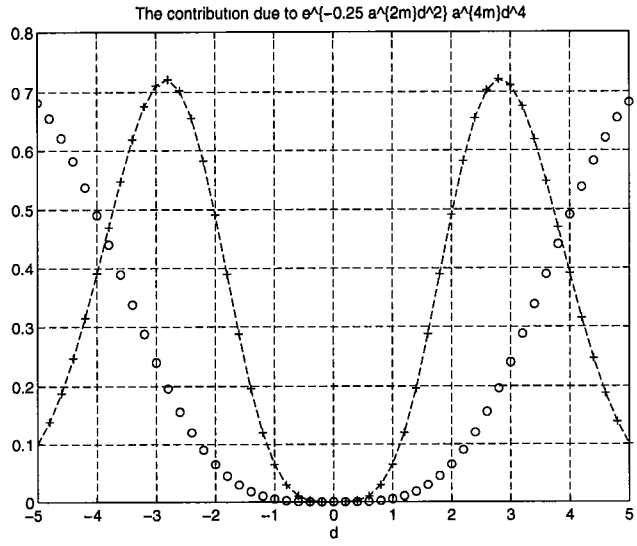


Figure A.1: $\cos(\alpha_i)$ in One Dimension

Then $b = 1.0493$ is a sensible choice, consistent with the conditions on frames.

Figure A.1 shows the effect of the last term in the above expression for $< \psi_{ij}, \psi_{nj} >$. The graphical form of $< \psi_{ij}, \psi_{nj} >$ is essentially similar to the Mexican hat function (mother wavelet) for small values of d (until the first zero-crossing point) and begins to differ from it for larger values of d.

More generally, when the dilation levels are also different, i.e., $\omega_i = a^{m_i}$ and $\omega_n = a^{m_n}$, a similar derivation results in

$$\cos(\alpha_{nj}) = \frac{4}{3}\sqrt{\frac{2\omega_j^2}{\omega_{ji}\omega_{jn}}}\frac{\omega_j^4}{\omega_{ji}^2\omega_{jn}^2}e^{-\frac{\omega_j^2 d_j^2}{2}}\left(d_j^4\omega_j^4 - 6\omega_j^2 d_j^2 + 3\right)$$

where $\omega_j = \frac{\omega_{ji}^2\omega_{jn}^2}{\omega_{ji}^2+\omega_{jn}^2}$.

A.4 Barbalat's Lemma

If $f(t)$ is a uniformly continuous function, such that $\lim_{t\to 0}\int_0^t f(\tau)d\tau$ exists and is finite, then $f(t) \to 0$ as $t \to \infty$.

Bibliography

[1] K. J. Astrom and B. Wittenmark. *Adaptive Control.* Addison-Wesley, New York, 1989.

[2] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Info. Theory*, 39:930–945, 1993.

[3] F. C. Chen and H. K. Khalil. Adaptive control of nonlinear systems using neural networks–a deadzone approach. In *Proceedings of the American Control Conference*, pages 667–672, 1991.

[4] S. Chen, C.F.N. Cowan, and P.M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Trans. on Neural Networks*, 2-2:302–391, 1991.

[5] C. K. Chui and X. L. Shi. On multi-frequency wavelet decompositions. In Larry L. Schumaker and Glenn Webb, editors, *Recent Advances in Wavelet Analysis*, volume 3 of *Wavelet Analysis and its applications,*, pages 155–189. Academic Press, 1994.

[6] J. J. Craig. *Adaptive Control of Mechanical Manipulators.* Addison-Wesley, 1988.

[7] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. on Information Theory*, 36(5):961–1005, September 1990.

[8] I. Daubechies. *Ten lectures on wavelets.* CBMS-NSF regional series in applied mathematics. SIAM, 1992.

[9] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. Technical report, Stanford University, Dept. of Statistics, 1993.

[10] Jay Farrell and Walter Baker. Learning control systems. In P. Antsaklis and K. Passino, editors, *Intelligent and Autonomous Control Systems.* Kluwer Academic, 1992.

[11] J. H. Friedman. Projection regression. *Journal of the American Statist. Association*, 76:817–823, 1976.

[12] J. H. Friedman. Classification and multiple regression through projection pursuit. Technical Report LCS012, Stanford University, Dept. of Statistics, Stanford, 1985.

[13] J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991.

[14] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory, and neural networks. Preprint, 1994. Accepted for publication on Neural Computation in May, 1994.

[15] T. Holcomb and M. Morari. Local training for radial basis function networks: Solving the hidden unit problem. In *The American Control Conference*, pages 2331–2336, Boston, 1991.

[16] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[17] V. Kadirkamanathan and M. Niranjan. A function estimation approach to sequential learning with neural networks. *Neural Computation*, 5:954–975, 1993.

[18] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. Technical Report 619, NewYork University, Dept. of Computer Science, August 1993.

[19] K.S. Narendra and K. Parthasarathy. Identification and control of dynamical systems using neural networks. *IEEE Trans. on Neural Networks*, 1:4–27, 1990.

[20] Kumpati S. Narendra. Adaptive control using neural networks. In W. T. Miller, R. S. Sutton, and P. J. Werbos, editors, *Neural Networks in Control.* MIT Press, Cambridge, MA, 1990.

[21] Kumpati S. Narendra and Snehasis Mukhopadhayay. Adaptive control of nonlinear multivariable systems using neural networks. *Neural Networks*, 7(5):737–752, 1994. Invited Article.

[22] P. Niyogi and F. Girosi. On the relationship between generalization error,hypothesis complexity, and sample complexity for radial basis functions. A.I.Memo, MIT., February 1994. ftp ai-publications/1994/AIM-1467.ps.Z.

[23] Y. C. Pati and P. S. Krishnaprasad. Analysis and synthesis of feedforward neural networks using discrete affine wavelet transformations. *IEEE Trans. on Neural Networks*, 4(1):73–85, January 1993.

[24] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the Annual Asilomar Conference on Signals Systems and Computers*, November 1993.

[25] Y.C. Pati and P.S. Krishnaprasad. Discrete affine wavelet transforms for analysis and synthesis feedforward neural networks. In R. Lippman, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems III*, pages 743–749, San Mateo, CA, 1990. Morgan Kaufmann, Publishers.

[26] Tomaso Poggio and Federico Girosi. Extensions of a theory of networks for approximation and learning: dimensionality reduction and clustering. MIT AI Memo No 1167, April 1990.

[27] Tomaso Poggio and Federico Girosi. Networks for approximation and learning. In *Proceedings of the IEEE*, volume 78, pages 1481–1497. IEEE, 1990.

[28] Marios M. Polycarpou and Petros A. Ioannou. Stable nonlinear system identifiaction using neural network models. In George A. Bekey and Kenneth Y. Goldberg, editors, *Neural Networks in Robotics*, pages 147–164. Kluwer Academic Publishers, 1993.

[29] T.D. Sanger. A tree-structured adaptive network for function ap proximation in high-dimensional spaces. *IEEE Trans. on Neural Networks*, 2(2):285–293, March 1991.

[30] R. Sanner and S. Slotine. Gaussian networks for direct adaptive control. *IEEE Trans. on Neural Networks*, November 1992.

[31] Shankar Sastry and Marc Bodson. *Adaptive Control: Stability, Convergence, and Robustness.* Prentice-Hall, 1989.

[32] B. Widrow and M. E. Hoff. Adaptive switching circuits. In *IRE WESCON*, convention record Part 4, pages 96–104. IRE, New York, 1960.

[33] B. Widrow and S. D. Stearns. *Adaptive Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, 1985.

[34] Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. on Neural Networks*, 3(6):889–898, November 1992.